

## AUTOMATIC STEMMING OF ALBANIAN THROUGH A RULE-BASED APPROACH

Jetmir Sadiku and Marenglen Biba

University of New York in Tirana, Albania

### Abstract

*Text Mining is a knowledge-intensive technique that is used to automatically process a collection of documents by employing a set of analysis tools. This growing field has increasingly become important nowadays in order to gather data in the form of unstructured text and infer information or knowledge from it. Many fields such as medicine, banking, finance, marketing, spam filtering etc. are largely benefiting from data mining techniques due to large repositories of text documents available. This paper deals with the design and implementation of a stemming algorithm for the Albanian language. A stemming algorithm is a procedure that removes the suffixes from the words providing the root (stem) of the words. Stemming is an important step in natural language processing and as a consequence in text mining. We exploit a rule-based approach to implement the algorithm and then the proposed algorithm is used to classify corpuses of text documents. Experimental results show the effectiveness of our approach.*

**Key words:** text mining, stemming, natural language processing, Albanian, document classification

## 1. INTRODUCTION

Today we live in a world governed by information. Information comes in different forms; from different fields of life and of course in different ways. It is said that we are living in the information age and large amounts of data are stored mostly in text format. This amount of data is available on enterprise intranets, on the Internet and elsewhere we can access the information in electronic way. In order to gain knowledge from information and then to use it for any purpose that we need it, we have to process this information.

Text Mining is a knowledge-intensive technique that is used to interact with a collection of documents by employing a set of analysis tools (Feldman and Sanger, 2007). Text mining is a subfield of Artificial Intelligence (AI) that is used to process the information in collaboration with other techniques from related subfields of AI. Data mining, knowledge management, machine learning, information retrieval (IR), Information Extraction (IE) and Natural Language Processing (NLP) are the techniques that together with Text Mining form the basis of processing and deriving knowledge from text.

Information Retrieval (IR) is the process of finding the document that contains answers to questions but it does not retrieve the answers itself (Manning et al., 2008). IR systems execute a set of steps in order to achieve their objective. Among different steps used by IR systems, this dissertation is interested only in the steps related with the pre-processing of the documents: deleting the stopwords and stemming the terms. These two steps are analyzed in depth throughout the dissertation. One important step is the stemming of terms. This step is both important and difficult because of the fact that it is language dependent and involves good knowledge in NLP and computational linguistics.

A stemming algorithm is a procedure that removes the suffixes from the words providing the root (stem) of the words. For example, the words player and playing have the same root play. The task of stemming algorithms is analogous with the number normalization that is done to achieve similarities. Different stemming algorithms which pre-process text in different languages like English, Spanish

and others have been created and improved during the past, but for the Albanian language such algorithms have not been done. This dissertation represents a first set of rules for Albanian language that will be used in a stemming algorithm. Further this dissertation discusses parts that are related with document preprocessing, mostly based on computational linguistics.

Natural Language Processing (NLP) is a technique which makes possible to have a better understanding of natural languages by use of computers (Kao and Poteet, 2006). Relevant studies on the morphology and syntax of Albanian are used during the dissertation for the purpose of creating the list of stop words and finding the stem of the words of Albanian.

In this paper we present the design and development of a stemming algorithm for the Albanian language in order to use it for text classification. This algorithm takes as input unstructured text in Albanian and for every word in the document it finds the stem of the word. The stems of all the words will then be put in a vector. After developing this algorithm, a group of different text documents will be used as input for the stemming algorithm and after that, an evaluation of different machine learning algorithms will be performed in order to show the effectiveness of the stemming algorithm produced.

## 2. RELATED WORK

To the best of the author's knowledge, there have not been any approaches yet in the implementation of an Albanian stemmer. This is due to the fact that not much research has been dedicated to this problem previously but also due to the difficulty of the Albanian language. Only a few papers on the Albanian language have been published, but no real studies for stemming have been performed. These papers deal mostly based with Albanian dictionaries without going much into the hardest part which is that of processing of Albanian.

The problem gets even more complicated because of the fact that today's Albanian Literature Language was done formal by the reform in 1972 unifying two dialects Gheg and Tosk while J. B. Lovins published the first paper per a stemming algorithm in 1968. In addition to that, political situations for some decade affected negatively the research in this area. There are also some other issues especially of technological nature that have had direct impact on the absence of the research for the field of stemming algorithms. Here we review a few works which are related to our research.

The work in (Lagji et al., 2007), presents an electronic dictionary for Albanian. The research was focused on building NLP tools for the Albanian language. However, their work was based on the usage of a database of words. Their purpose was to be able to automatically recognize words of a particular text as well as to recognize some morphological and syntactical features.

Another interesting study is done in (Trommer and Kallulli, 2003). Their study can be considered the first step towards the corpus linguistic for Albanian. They present a morphological tagger used as a main component in a part-of-speech tagging system for a corpus of standard Albanian. Albanian language is very rich in inflectional paradigms making the research and development very challenging. The authors analyze the main morphological components of the Albanian language and represent components of the tagger system more in detail.

As research on Albanian is in the early steps, this paper aims to give an important contribution in the field with the goal of developing a set of rules for stemming words in Albanian and a stopword list that will help in classifying text documents in Albanian.

## 3. ALBANIAN LANGUAGE MORPHOLOGY

Albanian is an Indo-European language that is spoken by nearly 7.6 million people not only in Albania and Kosovo but also in western Macedonia, southern Montenegro, southern Serbia and north-western Greece. Because of the emigration, speakers of Albanian can be found in Turkey, Germany, Switzerland, United Kingdom, Scandinavia, Netherlands, Australia, New Zealand, United States, Brazil and Canada (Wikimedia Foundation, 2011b). The linguistic material of this paper is based on

the book "*Gramatika e Gjuhes Shqipe*" (Grammar of the Albanian Language) edition of the Albanian Academy of Sciences, Institute of Language and Literature (Agalliu et al., 2002).

### 3.1 Morphological features

Morphology is that part of grammar which deals primarily with the study of pattern's formation as well as studying the meaning of these forms (Agalliu et al., 2002). Morphology studies words in two perspectives, as part of speech and as meaning forms. Morphology does not deal with individual words. It examines the common features of words of the same class, drawing rules of a general nature. Morphology has connections with phonetics as well as syntax. Studying these connections from the perspective of finding the root of the word does not constitute interest for the dissertation; hence the study will be based on terms of word formation and not in terms of their syntax and phonetics. All the words of a language are grouped in classes of lexicon-grammar categories which are called parts of speech. This classification is done based on joint study of the lexical and grammatical characteristics. Albanian clearly distinguishes these parts of speech: noun, adjective, pronoun, verb, adverb, preposition, connectors, particles and exclamatory.

This paper will focus extensively on the ways of forming these parts of speech in order to make possible to derive the general rules used to construct the algorithm for finding the stem of words. As stated in (Agalliu et al., 2002), the word is the basic unit of language. Usually the word in a language exists as a system of word forms. Most words can be broken down into smaller units which carry out lexical grammatical meanings (*qytet-ar*, *hekur-os*, *për-dredh* etc). There are also words which consist of a single morpheme (*afër*, *larg*, *drejt* etc). Mostly these kinds of morpheme are represented by the adverbs, prepositions and connectors. Of particular interest is the definition of the types of morpheme because based on these types we can understand how words are formed and how it will be possible to stem words. From the perspective of their grammatical understanding, morphemes can be divided into two groups: Stemming Morphemes and Affixing Morphemes.

Every word that can be divided, from the structural point of view, has at least one rooting morpheme. The root (stem) represents the lexical core of the word. Every word has a root, meaning that word without root does not exist. In the words with one morpheme, morpheme itself constitutes the root of the word. Affix morphemes serve to form other words, giving a new grammatical meaning. These morphemes are divided into prefixes, suffixes, endings and joints. Among them, the most important are the prefixes and suffixes. For each of the meanings of word formation, a table with the affix and some illustrative examples that supports them will be build. In this paper we will not extensively deal with the semantic, grammar or syntax analysis. After completing the study of the ways of forming various parts of speech, not all forms will be used in the rules of the algorithm. Prefix morphemes are called those which stand in front of the word's root or before another prefix (*nën-drejt*, *mos-besim*, *ri-përdor* etc.). Suffix is called the morpheme which lies behind the root of the word (*plag-os*, *hap-ur*, *rrjedh-im* etc.). The methods of constructing the stemming algorithm will depend on the rules that will emerge from the word formation analysis.

### 3.2 Nouns Formation

Noun is called the part of speech that names living beings and things. Noun as part of speech includes grammatical meanings of gender, the number and the race. In this paper we are concerned with the way words form the noun and not in other directions of syntax or phonetic. In this way we will rely only on those parts of interest to this paper. Nouns of today's Albanian language are added and constantly enriched with new words. Nouns in Albanian are mainly formed from adjectives, verbs and other nouns. The primary ways of forming words in the class of nouns are the origin and composition. Origin can be: suffix, prefix, suffix-prefix and affix-free. The first three consist of particular interest because they are the main sources of the birth of new forms, while the fourth has no great impact.

### 3.2.1 Nouns Formation with Prefixes

Prefix formation of nouns is not very productive. Nearly twenty prefixes can be used but a small number among them are productive. A part of the prefixes stem from adverbs and are used as a preposition. Prefixes *mos-* and *pa-* in many cases are placed in front of adverbial nouns of action giving them opposite sense. Prefixes which express adverbial meanings of time and space are numerous but are not so productive. The last group comes from foreign prefixes that are borrowed into Albanian. Table 1 shows them.

Prefixes	Examples
<i>mos-</i> , <i>pa-</i> , <i>para-</i> , <i>kundër-</i> , <i>sipër-</i> , <i>pas-</i> , <i>mbi-</i> , <i>ndaj-</i> , <i>për-</i> , <i>prej-</i> , <i>ultra-</i> , <i>a-</i> , <i>anti-</i> , <i>super-</i> , <i>auto-</i> , <i>poli-</i> , <i>bio-</i> , <i>gjeo-</i>	<i>mos-marveshje</i> , <i>mos-pajtim</i> , <i>pa-dije</i> , <i>pa-durim</i> , <i>para-ndjenjë</i> , <i>kundër-masë</i> , <i>kundër-thënie</i> , <i>sipër-faqe</i> , <i>sipër-marrës</i> , <i>pas-ardhës</i> , <i>pas-drekë</i> , <i>mbi-shkrim</i> , <i>mbi-vlerë</i> , <i>prej-ardhje</i> , <i>për-vojë</i> , <i>për-masë</i> , <i>ultra-tingull</i> , <i>a-simetri</i> , <i>a-ritmi</i> , <i>anti-fashist</i> , <i>anti-fetar</i> , <i>super-fuqi</i> , <i>super-prodhim</i> , <i>auto-didakt</i> , <i>poli-glot</i> , <i>poli-foni</i> , <i>poli-grafi</i> , <i>bio-kimist</i> , <i>gjeo-graf</i>

Table 1: Prefixes and illustrative examples

### 3.2.2 Nouns Formation with Suffixes

For the study of the nouns formation with the suffix, it should be noted that there is a connection in semantics between the formative theme and the noun derived from this word. There are situations when the link in semantic cannot be felt by speakers and in this case nouns are not analyzed as derived nouns. Suffixes are distinguished for wealth and variety of meanings. Nearly all suffixes and some illustrative examples are presented in the Table 2.

Suffix	Examples
-ës, -ar, -tar, -atar, -or, -tor, -ak, -as, -an, -it, -iot, -ist, -ik, -ant, -ent, -ier, -xhi, -xheshë	<i>mbledh-ës</i> , <i>shkel-ës</i> , <i>vjel-ës</i> , <i>argjend-ar</i> , <i>kepuc-ar</i> , <i>gjah-tar</i> , <i>shkrim-tar</i> , <i>këmbës-or</i> , <i>mirdit-or</i> , <i>parim-or</i> , <i>faj-tor</i> , <i>mina-tor</i> , <i>pune-tor</i> , <i>fier-ak</i> , <i>durrs-ak</i> , <i>dibr-an</i> , <i>kuksi-an</i> , <i>kavaj-as</i> , <i>tiran-as</i> , <i>gjirokastr-it</i> , <i>mallakastr-iot</i> , <i>kitar-ist</i> , <i>makin-ist</i> , <i>nevrastr-ik</i> , <i>kurs-ant</i> , <i>muzik-ant</i> , <i>asist-ent</i> , <i>bank-ier</i> , <i>port-ier</i> , <i>qira-xhi</i> , <i>qira-xheshë</i>
-je, -ishte, -urina, -(ë)sirë	<i>mbath-je</i> , <i>vesh-je</i> , <i>bar-ishte</i> , <i>far-ishte</i> , <i>mbet-urina</i> , <i>qelq-urina</i> , <i>plaçk-urina</i> , <i>kalb-ësirë</i> , <i>ëmbël-sirë</i>
-ishte, -ore, -tore, -inë, -ajë, -(ë)tirë, -(ë)sirë	<i>fidan-ishte</i> , <i>lul-ishte</i> , <i>pem-ishte</i> , <i>mëngjes-ore</i> , <i>krip-ore</i> , <i>gjell-tore</i> , <i>ëmbël-tore</i> , <i>kodr-inë</i> , <i>cmend-inë</i> , <i>shkret-ëtirë</i> , <i>zbraz-ëtirë</i> , <i>hap-ësirë</i>
-(ë)si, -(ë)ri, -shëm, -shmëri, -im, -je, -esë, -atë, -imë, -më, -llëk, -izëm, -azh, -urë, -ikë	<i>gjat-ësi</i> , <i>mbar-ësi</i> , <i>vazhdim-ësi</i> , <i>bashk-ësi</i> , <i>djal-ëri</i> , <i>fisnik-ëri</i> , <i>gati-shmëri</i> , <i>pjell-shmëri</i> , <i>bes-im</i> , <i>drejt-im</i> , <i>vend-im</i> , <i>njoh-je</i> , <i>marr-je</i> , <i>jet-esë</i> , <i>blegëri-më</i> , <i>bubulli-më</i> , <i>avokat-llëk</i> , <i>budalla-llëk</i> , <i>kapital-izëm</i> , <i>parashut-izëm</i> , <i>spiun-azh</i> , <i>agjent-urë</i> , <i>metod-ikë</i> , <i>estet-ikë</i>
-ac, -acak, -acuk, -aq, -alaq, -aluq, -alec, -arash, -avec	<i>morr-acak</i> , <i>rren-acak</i> , <i>vjedh-acak</i> , <i>rrjep-acuk</i> , <i>verdh-acuk</i> , <i>trash-aluq</i> , <i>shtremb-aluq</i> , <i>bark-alec</i> , <i>grind-avec</i> , <i>qull-avec</i>

Table 2: Suffixes and illustrative examples

### 3.3 Composite Formation of Nouns

Suffix formation of nouns is the most productive way and after that comes the composition way. On the basis of syntactic relations between the limbs we distinguish between composites with conjunctive relations and composites with subordinate relations. The latter one is more productive. We distinguish various types of composite subordinates:

Words composed of two nouns

1. Noun + adverbial noun of actors, formed with the suffix *-ës*  
(*bukë + pjek-ës*, *gur + skalit-ës*, *rroba + qep-ës* etc)
2. Noun + adverbial action noun, formed with the suffixes *-je* or *-im*  
(*gjak + derdh-je*, *besë + lidh-je*, *letër + këmb-im*, *dëm + shpërbl-im* etc)
3. Noun + whatever noun  
(*ditë + lindje*, *mes + ditë*, *vend + banim*, *vaj + guri*, *qymyr + guri* etc)

Words consisting of an adverb and a noun

(*bashkë+bisedim*, *bashkë+luftëtar*, *drejt+shkrim*, *keq+kuptim* etc)

Words consisting of a pronoun and a noun

(*vetë+vendosje*, *vetë+sherbim*, *dy+luftim*, *tre+mujor* etc)

Composites formed by a noun and an adjective

(*gushë+kuq*, *lulë+kuqe*, *gojë+mberthyeri* etc)

Composites formed by a verb and a noun

(*thith+lopë*, *frymë+marrje* etc)

All above cases will be treated as special situations of finding the root of the word and special attention will be paid to them.

### 3.3 Adjectives Formation

Adjective is called that part of speech that names a feature of the item and fits in gender, in number and for a good part of the cases with the name of this thing (Agalliu et al., 2002). Because adjectives name a feature of a thing in the sentence, we can say that from a grammatical point of view, nouns and adjectives are related. Words that are part of the adjectives group represent a variety of forms different from the nouns group.

The adjective as part of the speech is enriched over time and continues to be enriched with new words. Productive ways of word formation of adjectives are (same as the noun) origin, conversion and composition. In the class of adjectives we recognize a specific class of adjectives which includes those adjectives that cannot be analyzed in their component parts hence neither forming words nor themes, nor can any other form of word building them be understood.

This particular group also includes the adjectives that are derived from foreign languages (*absolut* in Albanian is *absolute* in English; *banal* in Albanian is *banal* in English etc.). As well as nouns, adjectives formation includes word formation as adding prefixes, suffixes, nodes and nodes and suffixes together. These classes will be studied below.

#### 3.3.1 Adjectives Formation with Prefixes

This way of forming words is regarded as somewhat productive way to create adjectives. This is not from the large number of prefixes but because some of them are very productive. Most productive prefixes are *pa-* and *jo-*. Prefix *jo-* is added to adjectives usually formed with suffix *-shëm*. Similar to prefix *pa-* is the prefix *mos-* but it is less productive. Table 3 shows them.

Prefixes	Examples
pa-, jo-, mos-, anti-, para-, pro-, pas-, prapa-, stër-, mbi-, ndër-, bashkë-, nën-, pan-, tej-, sapo-, ç-, sh-, zh-, i-, in-, an-, super-, lart-, poshtë-, sipër-, drejt-, gjysmë-, gjithë-, jashtë-, plot-, shumë-, vetë-, filo-, gjeo-, hidro-, neo-, mono-, poli-, pseudo-	i pa-afrueshëm, i pa-botuar, jo-fetar, jo-normal, mos-mirënjohës, mos-përfillës, anti-fashist, anti-fetar, para-shkollor, para-ushtarak, pas-universitar, prapa-mbetur, prapa-vendosur, pro-amerikan, pro-revizjonist, i stër-madh, i stër-gjatë, mbi-tokësor, mbi-njerëzor, ndër-luftues, ndër-sektorial, bashkë-kohor, bashkë-përgjegjës, nën-ligjor, nën-ujor, pan-amerikan, pan-aziatik, i tej-dukshëm, i tej-ngopur, i sapo-formuar, i sapo-lindur, ç-njerëor, i ç-regullt, i sh-kujdesur, i-moral, in-organik, an-alfabet, i lart-përmendur, i lart-shënuar, i poshtë-shënuar, gjysmë-zyrtar, gjysmë-shekullor, gjithë-fuqishëm, gjithë-pushtetshëm, jashtë-fisnor, i jashtë-zakonshtëm, i plot-fuqishëm, i plot-pushtetshëm, shumë-shekullor, shumë-vjear, vetë-kritik, vetë-mohues, filo-grek, filo-rus, gjeo-elektrik, gjeo-magnetik, hidro-elektrik, hidro-energjitik, mono-silabik, neo-kolonial, neo-latin, poli-teknik, poli-fonik

Table 3: Prefixes and illustrative examples

### 3.3.2 Adjectives Formation with Suffixes

The method of forming adjectives with suffixes is one of the productive ways with which we can form adjectives. Nouns, adjectives, numerals, verbs and adverbs can serve as formative themes for the forming adjectives with the suffixes. Table 3.4 shows them.

Suffixes	Examples
-ës(e), -ar, -tar, -ues, -or, -tor, -ak, -ësh, -ac, -anik, -ac, -acak, -alak, -ik, -al, -ual, -iv, -shëm, -ueshëm, -të, -ë -ur, -uar	djeg-ës(e), mbledh-ës(e), bregdet-ar, mesjet-ar, planet-ar, mesdhe-tar, përfundim-tar, bim-or, trup-or, vet-or, bari-tor, lesh-tor, perandor-ak, vez-ak, frik-acak, mburr-acak, agronom-ik, akadem-ik, atom-ik, embrion-al, eksperiment-al, objekt-iv, edukat-iv, i buj-shëm, i domosdo-shëm, i mrekull-ueshëm, i trisht-ueshëm, i shpesh-të, i rëndom-të, i dredh-ur, i hap-ur, i vendos-ur, i kupt-uar, i afr-uar

Table 4: Prefixes and illustrative examples

### 3.3.3 Composite Formation of Adjectives



Compound adjectives are formed from merging two (or three) themes in a single word. Like the formation of composite nouns, we also distinguish two types of adjectives:

- a) Composites with conjunctive relations
- b) Composites with subordinate relations

The first group contains those kinds of adjectives that do not have nodes and that are equivalent from both the semantic and syntactic (*anglo-amerikan, agraro-industrial, meterialo-teknik* etc.). In the group of subordinate relations we distinguish these subgroups:

1. Composed adjectives with two nominal themes  
(*hundë + shkabë, zemër + lepur, zemër + gur, kokë + kungull* etc.)
2. Composed adjectives with a nominal theme and a adjective theme  
(*bark + gjerë, bel + hollë, kokë + prerë* etc.)
3. Composed adjectives with two adjective themes  
(*elektro + magnetik, elektro + mekanik, gjermano + lindor* etc.)
4. Composed adjectives with a numeral and an adjective  
(*një + mujor, treqind + vjekar, shumë + vje\_car, disa + ditor* etc.)
5. Composed adjectives with a partial theme and an adverbial theme  
(*dasha + mir, dasha + keq* etc.)

### 3.4 The Verb as Part of the Speech

Verb is that part of speech that names an action as a process and that has category of number, manner and time (Agalliu et al., 2002). Most of the verbs name actions that are performed or are suffered by the subject (*edukohem, punoj, regjistrohem* etc.). Another group of verbs name the state of the subject (*jam, gjendem* etc.) or changes of the state of the subject (*rritem, skuqem, zverdhem* etc.). Based on their origin, verbs of the Albanian language can be frontal, derivative and composed. Frontal verbs are those which are not formed from other words or with the help of affixes. This group of verbs contains:

- a) *La-j, mba-j, tha-j, gri-j, kua-j*
- b) *Hap, mbyll, prek, qesh, qaj, dal*
- c) *Irregular Verbs*

It seems clear that as frontal verbs we can take those verbs whose themes of today's Albanian language are not analyzed. Derived verbs are formed mainly from the nouns, adjectives, adverbs or other verb. Formation from suffixes is more productive in the group of derived verbs. After that comes the formation from prefixes, with prefixes and suffixes at the same time and without any means of word formation.

#### 3.4.1 Verbs Formation with Suffixes

This group consists of two types of suffixes:

1. Suffixes of the derived verbs of the first conjugate

## 2. Suffixes of the derived verbs of the second conjugate

Tables 5 and 6 show some examples.

## 3.4.2 Verbs Formation with Prefixes

The most productive prefixes for verb formation are shown in Table 7. In the group of verbs formed by prefixes and from other parts of the speech, the similar phenomenon like in formation without affixes is noticed.

Suffixes	Examples
-o-, -llo-, -(ë)ro-, -(ë)so-, -(ë)to-, -ë)zo-	nder-o-j, pun-o-j, ndrysh-o-j, turbu-llo-j, the-llo-j, vezu-llo-n, flak-ëro-j, mish-ëro-j, lajm-ëro-j, formë-so-j, vlerë-so-j, frikë-so-j, ledha-to-j, dëm-to-j, viza-to-j, pyll-ëzo-j, shtet-ëzo-j, udhë-zo-j

Table 5: Suffixes of the derived verbs of the first conjugate

Suffixes	Examples
-os, -s/is, -it	burg-os, bring-os, damk-os, arrati-s, bezdi-s, lemeri-s, rast-is, tigan-is, vaj-is, gjob-it, darov-it, ngul-it, shkoq-it

Table 6: Suffixes of the derived verbs of the second conjugate

Prefixes	Examples
për-, sh-, ç-, zh-, z/s-, n/m-, mbi-, nën-, ndër-, ri-, stër-, de-	për-mbyt, për-hap, për-caktoj, sh-faq, sh-koq, sh-pik, ç-mallem, ç-armatos, ç-rregulloj, zh-vesh, zh-vlerësoj, z-gjat, z-buloj, s-kuq, s-pastroj, mbi-ngarkoj, mbi-kqyr, nën-shkruaj, nën-kuptoj, ndër-marr, ndër-lidh, ri-zgjedh, ri-organizoj, ri-prodhoj, stër-holloj, stër-mundoj, de-maskoj, de-gradoj, de-shifroj

Table 7: Prefixes and illustrative examples

## 3.4.3 Formation with Prefixes and Suffixes at the Same Time

A part of the derivative verbs are formed simultaneously from the prefix and suffix theme, which in most cases are with nominal nature and rarely with adjective nature.

1. Verbs of the type *për* + formative theme + suffix

- *për* + nominal theme + -o-  
(*për-fund-o-j*, *për-gënjeshtr-o-j*, *për-mall-o-j* etc.)
- *për* + nominal theme (pronoun or adjective) + -so-  
(*për-faqë-so-j*, *për-gjithë-so-j*, *për-mirë-so-j* etj)

2. Verbs of the type *sh-(zh-)* + formative theme + suffix



- *sh-* + nominal theme + *-ëzo* (*sh-fryt-ëzo-j*)
- *sh-/-* + nominal theme + *-so/-o-* (*sh-pronë-so-j*, *sh-pët-o-j*)
- *zh-* + nominal theme + *-o/-ëso-* (*zh-dogan-o-j*)

3. *z-* + adjective theme + *-o-* (*z-bukur-o-j*, *z-gjer-o-j*)

#### 3.4.4 Composite Formation of Adjectives

Verbs as composed words are scarce in the Albanian language. Among them the main determinants are composite of the type verb + adverb (*mirë + kuptoj*, *mirë + mbaj*, *keq + trajtoj* etc.). There are verbs in which the first limb determines the action expressed by the second limb (*buzë-qesh*, *duar-trokas*, *rreth-shkruaj* etc.). Also there are cases when before the verb an adverb stands (*bashk + bisedoj*, *para + caktoj*, *para + paguaj vetë + mbrohem* etc.). All these cases are considered as special cases and need a special treatment that in most cases goes beyond the research of this paper.

#### 3.5 Adverbs Formation

Adverbs name a feature of an action or of the situation, the circumstances in which this action is certified or shows a level of quality of circumstance or the intensity of an action (Agalliu et al., 2002). In contrast to other parts of the speech, adverb is not variable. Adverbs as part of the speech have not fixed grammatical categories. It has only a category of the scale, which is expressed only in an analytical way (like adjectives). Other features of adverbs are:

- Formative and semantic connection with other parts of speech
- A special system of affixes in the case of the derived adverbs
- The large number of idioms

Adverbs are used regularly close to a verb, close to an adjective or close to other adverbs. Rarely may they be used close to a noun. Because adverbs are meaningful nominators, they are used as autonomous limbs in the sentence and that is why they are distinctive from the prepositions, connectors, exclamatory or particles. From the point of view of their meaning, adverbs have diversity but this does not imply much interest in our study. From the point of view of their formative ways, adverbs can be simple, derivative, compound and idiomatic.

##### 3.5.1 Simple Adverbs

This group includes all those adverbs that are not formed from other parts of the speech and from the perspective of today's Albanian language cannot be analyzed based on their meaning. Such adverbs are considered *afër* (near), *larg* (far), *keq* (worse), *mire* (well), *pas* (then), *prapa* (behind) etc.

##### 3.5.2 Derived Adverbs

This group includes those adverbs which are formed with the help of other parts of speech and by adding different affixes. The methods of suffix formation are more productive. Several subgroups in the group of derived adverbs can be distinguished:

1. Adverbs formed through conversion from other parts of speech
  - Nouns modified to be used as adverbs

Some adverbs are formed from different noun's categories without the help of other linguistic tools. By modifying the noun we change the meaning of the word thus giving birth of adverbs. Nouns used as adverbs are divided into several categories:

(a) Words in the derivative or objective form like *motit*, *sheshit*, *rrafsh*, *rresht*, *rreth*, *rrotull* etc. This group also includes words with naming origin like *fare*, *herët*, *krejt* etc.

(b) Another group consists of words like *ditën*, *natën*, *anash*, *një ditë*, *një natë*, *një mëngjez*, *një verë* etc. In this group the process of modification of nouns is not consolidated yet, thus in several cases the words can be taken as naming forms also.

(c) Another group consists of words that are related with the unchangeable features of the derivative form. For example, *së afërmi*, *së jashtmi*, *së mbari*, *së larti*, *së gjati*, *së tepërmi*, *së brendshmi* etc.

(d) A special group consists of nouns which only in certain context are used as adverbs. Those words are mostly nouns of the nominal race and rarely of the objective race (*copë*, *faqe*, *uturim*, *grumbull*, *pallë*, *tapë*, *thikë*, *varg* etc.).

- Verbs modified to be used as adverbs.

Adverbs without formative tools are also formed by different verbal forms. For example, *kaluar*, *ndyrë*, *hidhur*, *hapur*, *rrëmbyer*, *shtruar* etc. Modification of verbs to give birth of adverbs walks alongside the modification of nouns.

## 2. Adverb formed with the suffixes.

Suffix type formation is very productive, the same as in other parts of the speech. The most productive adverbial suffixes are: - (*i*)*sht*, -*as/azi* and -*thi*. Recently it is noticed that the Albanian literary language productivity of suffix -*shëm* is increasing. Table 8 represents suffixes and some illustrative examples.

Suffixes	Examples
-(i)sht, -as/azi, -thi, -shëm	detyrim-isht, ushtarak-isht, egërsi-sht, qetësi-sht, imtësi-sht, bark-as, bark-azi, radh-as, radh-azi, djatht-as, djatht-azi, kalim-thi, vetëtim-thi, fluturim-thi, furi-shëm, hare-shëm, natyr-shëm

Table 8: Suffixes and illustrative examples

### 3.6 Preposition, Conjunction, Particle and Exclamatory

These parts of speech are studied in order to produce a stopword list needed for the stemming algorithm. Usually these parts of speech are characterized by their immutability. This is why they are treated as stopword list. For each of these parts of speech a table with words constituting each of them is presented as follows:

Prepositions
me, prej, nga, për, rreth, në, mes, bri, rrëzë, krye, falë, ndaj, nën, nëpër, ndër, sipas, mbi, më, pa, për arsye, për shkak, me anë, me anën, në drejtim, në sajë, në vend, kundrejt, prapa, përtej, afër, larg, te, tek, përve, pas, simbas, sipas, mbas, anëmbanë, brënda, drejt, gjatë, ve, tutje, rrotull, pranë, para, përballë, sipër, matanë, anë e përqark, rreth e rrotull, rreth e përqark, tej e tej, mes për mes, ballë për ballë, tok me, bashkë me, nëpërmjet, nëpërmes, gjithë

Table 9: Prepositions

Conjunctions
dhe, sepse, kur, edhe, në qoftë se, e, o, as, po, nga, me, për, prej, apo, ose, në, ku, kur, sa, se, si, tek, qoftë, ve, por, da, sesa, teksa, kurse, porse, nëse, porsì, sikurse, mbasi, ngaqë, meqë, ngase, ndërsa, përvese, vese, derisa, megjithatë, atë, këtë, jo, ashtu, ndonëse, qysh, do, le, mos, ndaj, prandaj, pra, ashtu si, megjithëse

Table 10: Conjunctions

Particle
pikërisht, pa, se, jo, ndoshta, desh, nuk, nja, de, dot, a, bile, vallë, po, sidomos, le, mos, s, as, pale, posi, si, thujajse, pothuajse, kushedi, që, as që, ja që, jo që, jo, po se po, se mos, vetëm që, jose jo, jo e jo, e, ë, dhe, edhe, pra, sa, se, sikur, gati, plot, rreth, vetëm, ja, deri, qysh, mu, madje, veanërisht, afërsisht, ëhë, mosni, mbase, para,

Table 11: Particle

Exclamatory
ua, bobo, shët, oburra, oburrani, qyqja, of of, pthu, ptu, he he, bërr, uf uf, oh oh, bubu, aha, oho, uhu, eu, ah, oj, uh, u, hopa, na, moj, hop, hë, sus, bah, piis, kuku, vuu, bam

Table 12: Exclamatory

#### 4. ALGORITHM ANALYSIS AND DESIGN

This section presents the rules that derive from the analysis in Section 3. A set of rules and the stop words list will be integrated in a single algorithm. After expressing what rules do, then they will be implemented in Java. In Section 3 we presented the ways of forming words in Albanian. The primary ways were affix formation and as special cases we considered the composed way of formation.

The algorithm has a set of rules that are examined one by one in the given order and one of the rules inside the set is allowed to be executed. Longest possible suffixes and prefixes are removed first. Another restriction is that a stemming rule is not applied if the word after stemming contains only one letter. Only stems with two or more letters exist in Albanian. A rule consists of one or more if conditions that depending on the case they are used, checks the length of the word, the starting or the ending of the word. Based on the fulfilled conditions, the stem is returned. Sometimes, there are words that contain two vowels or suffixes one after the other. For these cases, rules of suffix removal and vowel removal will be checked twice. An example of a rule is shown in Figure 1. The algorithm contains in total 134 rules.

When removing the suffix or prefix, the stem does not have a linguistic meaning because they are used as an index for the database of documents and the user is not presented with the stem list. Different from other languages, the Albanian language does not have a general rule for forming the plural so it is not possible to have a step for this situation. The same can be stated about the feminine, masculine and neutral gender. It is very difficult in Albanian to have a general rule for the feminine, masculine or

neutral gender. For that reason there is no dedicated step in the algorithm that deals with these situations. A special step in the algorithm will be the case of suffixes that end with vowel. Words that end with these suffixes will be processed before the step of vowel reduction. Figure 4 represents the sequence of steps that will be executed by the algorithm. Each step has its own rules, derived by the morphological analysis of Albanian.

```
// *****Rule 104*****//
if ({word.endsWith("ësh") || word.endsWith("esh")} && word.length() > 5) {
  if (word.contains("shpesh")) {
    return word;
  } else {
    word = word.substring(0, lastPos - 3);
  }
}
```

Figure 1: Rule example

#### Step 1: Prefix Reduction

In Albanian, the most productive way of forming parts of speech are prefixes and suffixes. The first step of the algorithm deals with prefixes. Rules used in this step derive from the morphological analysis of the previous chapter and contain the most important prefixes. A small number of them are not included in the algorithm because of the fact that they are not very productive and only few words are created with them.

#### Step 2: Suffixes that end with Vowel Reduction

Among all suffixes, there are some suffixes that end with a vowel like *-je* (*mbathje*, *veshje* etc.) or *-shmëri* (*ngjashmëri*, *gatishmëri* etc.) that are processed by this step. This step is executed before vowel removal step because if the last vowel is first, the whole suffix is not recognized by the suffix removal step.

#### Step 3: Vowel Reduction

There are a lot of words that are not affected by the first rule or are affected by it and give a form of the word that can be further improved by removing the ending vowel (*-a*, *-e*, *-ë*, *-i*, *-o*, *-u*), for example *mbivlerë* is modified by the first rule as *vlerë* but it still can be modified by removing the vowel. Words like *dera*, *hëna*, *fletore*, *punëtore*, *mësuese*, *djalë*, *vezë*, *libri*, *teli*, *djalo*, *biro* etc. that are not modified by the first rule can also be modified by this step.

#### Step 4: Suffix Reduction

As mentioned in the first step, suffix formation is especially the most important way of forming parts of speech in Albanian. Suffix removal step is divided into two steps because there are some suffixes that end with vowel and they all are considered in step 2.

#### Step 5: Vowel Reduction

After performing step 4, there are also some words that contain a vowel in the end, and it is better to remove those vowels. As an example, consider the word *gatishmëri* that from step 3 is converted as *gati*. This word can be processed further by removing the vowel *i* giving the stem *gat*.

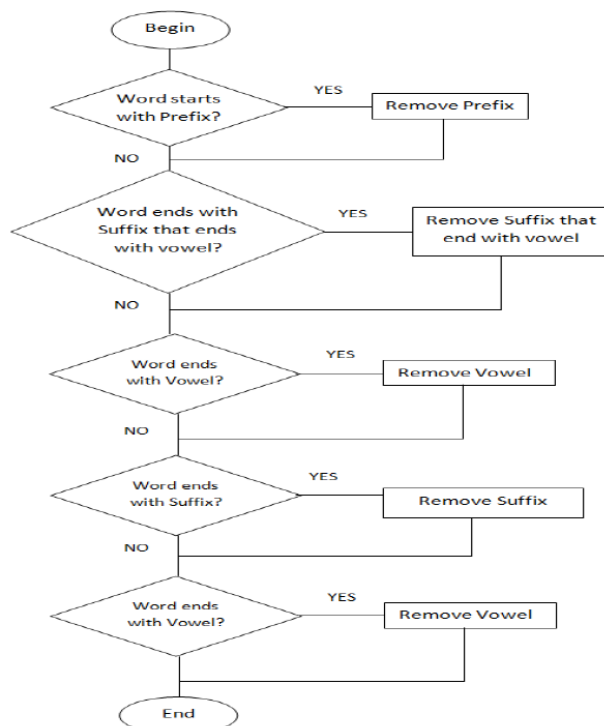


Figure 2: Sequence of steps of the algorithm

## 5. Experimental Evaluation

### 5.1 Experimental Setting

Usually, the effectiveness of stemming algorithms has been measured in terms of the impact that they give on the retrieval performance on a test collection. In this paper, we evaluate the algorithm in a text mining task. Tests are performed with the Weka software, which is an open source program that can be downloaded at <http://www.cs.waikato.ac.nz/ml/weka/>. Before running tests, we prepared a corpus of documents that have to be stemmed by our algorithm (JStem) and then classified by Weka.

Our corpora contain four main fields of study (Biology, History, Literary and Chemistry) with 40 documents each. Documents consist of text expressing some phenomenon or explaining some situation. Each document is firstly stemmed by the JStem algorithm and then it is modified by Weka constructing the input file *test.arff*. This file is then used to classify the documents and gives the performance of the classifying algorithm. The input file in Weka is modified by a filter that replaces the missing attributes and then by a StringtoWordVector filter.

The Data Mining algorithm used in these tests is Support Vector Machines (SMO). These algorithms are the most useful ones while we are mining data that is in numeric format. Tests are done between two fields that are not related with each other and with two fields that are related in order to see how accurate the algorithm is.

### 5.2 Experimental Results

In this section we present the experimental results on the task of text classification.

#### 5.2.1 Test 1: Classifying Biology and History Documents

After running the algorithm we see that the data mining algorithm is performing well (95% of instances is correctly classified) as shown in Figure 3. This high percentage is dedicated to the data mining and to the correctness of the stemming algorithm also. There are also a lot of other parameters that are shown by Weka. We will not focus on their meaning because it is out of our interest.

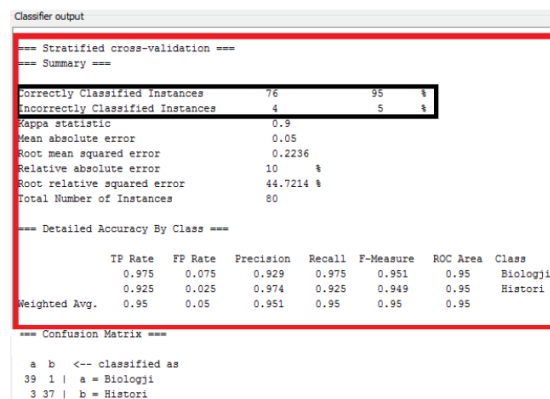


Figure 3: Results of mining with SMO for Test 1

As it is shown in Figure 3, the stemming algorithm is working well while it is used against two fields that are not related to each other. It is sure that this correctness will not be the same while we classify fields that are related with each other.

Now let us run the same tests with not stemmed documents. Figure 4 represents the results of mining of not stemmed documents. The performance of correctly classified documents without using the stemming algorithm is lower than in the case when we use the algorithm (for SMO function). The difference in performance is not high because the number and the content of the documents is not large.

```

Correctly Classified Instances      75      93.75 %
Incorrectly Classified Instances     5       6.25 %

```

Figure 4: Results of mining not stemmed documents with SMO

### 5.2.2 Test 2: Classifying Chemistry and Biology Documents

This is the situation where the classifying algorithms do not perform very well. This happens because in the text of biology and chemistry documents, at somewhat level words are the same, so the same will be the words produced by the stemming algorithm. There are some improvements that can be done in future in the stemming algorithm. Results of the test are presented in Figure 5.



```

Classifier output
Time taken to build model: 0.05seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      57      71.25 %
Incorrectly Classified Instances    23      28.75 %
Kappa statistic                    0.425
Mean absolute error                0.2875
Root mean squared error            0.5362
Relative absolute error            57.5 %
Root relative squared error        107.2381 %
Total Number of Instances         80

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
               0.675   0.25    0.73    0.675   0.701    0.713   Biologji
               0.75    0.325   0.698   0.75    0.723   0.713   Kimi
Weighted Avg.   0.713   0.288   0.714   0.713   0.712   0.713

=== Confusion Matrix ===
  a b  <-- classified as
 27 13 | a = Biologji
 10 30 | b = Kimi

```

Figure 5: Results of mining with SMO

Figure 6 presents the results of the mining of not stemmed documents. The performance of correctly classified documents without using the stemming algorithm is lower than in the case when we use the algorithm (for SMO function). The difference in performance in this case is high (from 71.25% to 66.25%).

```

Correctly Classified Instances      53      66.25 %
Incorrectly Classified Instances    27      33.75 %

```

Figure 6: Results of mining not stemmed documents

### 5.2.3 Test 3: Classifying History, Literary, Chemistry and Biology Documents

The last test is executed against four fields. As in all tests executed before, results show that the stemming algorithm represented in this dissertation performs well in fields that are not related but it needs improvements for fields that are related with each other. Figure 7 presents the results of mining the stemmed documents while Figure 8 presents the results of mining not stemmed documents. As we can see, the text classification algorithm is more accurate in the case of classification of stemmed documents and this demonstrates again the effectiveness of our approach.

```

Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      133     83.125 %
Incorrectly Classified Instances    27     16.875 %
Kappa statistic                    0.775
Mean absolute error                0.2682
Root mean squared error            0.3393
Relative absolute error            71.5278 %
Root relative squared error        78.3511 %
Total Number of Instances         160

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
               0.7     0.075   0.757    0.7     0.727    0.9     Biologji
               0.875   0.008   0.972    0.875   0.921    0.957   Histori
               0.775   0.108   0.705    0.775   0.738    0.887   Kimi
               0.975   0.033   0.907    0.975   0.94     0.969   Letersi
Weighted Avg.   0.831   0.056   0.835   0.831   0.832    0.928

=== Confusion Matrix ===
  a b c d  <-- classified as
 28 1 11 0 | a = Biologji
 1 35 1 3 | b = Histori
 8 0 31 1 | c = Kimi
 0 0 1 39 | d = Letersi

```

Figure 7: Results of mining with four classes

Correctly Classified Instances	126	78.75	%
Incorrectly Classified Instances	34	21.25	%

Figure 8: Results of mining not stemmed documents with SMO for Test 5

## 6. CONCLUSIONS AND FUTURE WORK

This paper presents a stemming algorithm for the Albanian language. A stemming algorithm is a procedure that removes the suffixes from the words providing the root (stem) of the words. Stemming is an important step in natural language processing and as a consequence in text mining. We exploit a rule-based approach to implement the algorithm and then the proposed algorithm is used to classify corpuses of text documents. We show through experiments on text classification that the classifying algorithm is more accurate when documents to be classified are first stemmed.

As future work we intend to develop the algorithm further. First, part-of-speech tagging is needed to overcome difficulties on stemming the compound ways of formation of different parts of the speech of Albanian. Another direction of research is dealing with rules for the compound formation which is this paper we have not considered. One possible approach for this is to have a database that contains all compound words of Albanian. When the algorithm finds in the text document a word that is composed of other words, and finds it in this database, predefined rules help to find the best stem or it may happen that two stems can be derived from that word.

## REFERENCES

- F. Agalliu, E. Angoni, Sh. Demiraj, A. Dhrimo, E. Hysa, E. Lafe, and E. Likaj. Gramatika e Gjuhes Shqipe. Akademia e Shkencave, 2002.
- J. Dawson. Suffix removal and word conation. ALLC Bulletin, 2:33:46, 1974.
- R. Feldman and J. Sanger. The Text Mining Handbook. Cambridge University Press: Cambridge CB2 8RU, UK, 2007.
- N. Ferro. Cross language evaluation forum web site, 2007. URL <http://www.clef-campaign.org/>.
- N. Ferro, M. Melucci, G. M. Di Nunzio, and N. Orio. The University of Padova at clef 2003: Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. Working Notes for the CLEF 2003 Workshop, pages 211-223, 2003. URL [www.clef-campaign.org/2003/WN\\_web/27.pdf](http://www.clef-campaign.org/2003/WN_web/27.pdf).
- P. Funchun, A. Nawaaz, L. Xin, and L. Yumao. Context Sensitive Stemming for Web Search. Annual International ACM SIGIR conference on Research and development in information retrieval, 30:639-646, 2007. 12
- A. Kao and S. R. Poteet. Natural Language Processing and Text Mining. Springer, 2006.
- R. Krovetz. Viewing Morphology as an Inference Process. Annual International ACM SIGIR conference on Research and development in information retrieval, 16:191-202, 1993.

K. Lagji, O. Piton, and R. Pernaska. Electronic dictionaries and transducers for automatic processing of the albanian language. International Conference on Applications of Natural Language to Information Systems, 12:407-413, 2007.

E. D. Liddy. How a search engine works, 2001. URL <http://www.cnlp.org/publications/02HowASearchEngineWorks.pdf>. Web of Deception 10/28/01.

J. B. Lovins. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11:22{31, 1968.

Ch. D. Manning, P. Raghavan, and H. Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.

J. May\_eld and P. McNamee. Single n-gram stemming. Annual International ACM SIGIR conference on Research and development in information retrieval, 26:415-416, 2003.

T. M. Mitchell. Machine Learning. McGraw-Hill, 1997.

R. Mitkov. The Oxford Handbook of Computational Linguistics. Oxford Univeristy Press, 2004.

Ch. D. Paice. An evaluation method for stemming algorithms. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 17:42-50, 1994.

F. Porter. Stemming algorithms for various european languages, 2006. URL <http://www.snowball.tartarus.org/texts/stemmersoverview.html>.

J. Trommer and D. Kallulli. A morphological tagger for standard albanian, 2003.

Inc. Wikimedia Foundation. Proto-indo-european language, 2011a. URL [http://en.wikipedia.org/wiki/Proto-Indo-European\\_language](http://en.wikipedia.org/wiki/Proto-Indo-European_language). 32

Inc. Wikimedia Foundation. Albanian language, 2011b. URL [http://en.wikipedia.org/wiki/Albanian\\_language](http://en.wikipedia.org/wiki/Albanian_language). 16, 32

P. Willett. The porter stemming algorithm: Then and now. Program: Electronic Library and Information Systems, 40 (3):219{223, 2006. 8

J. Xu and W. B. Croft. Corpus-based stemming using co-occurrence of word variants. ACM Transactions on Information Systems, 16:61-81, 1998.