Stemming algorithms

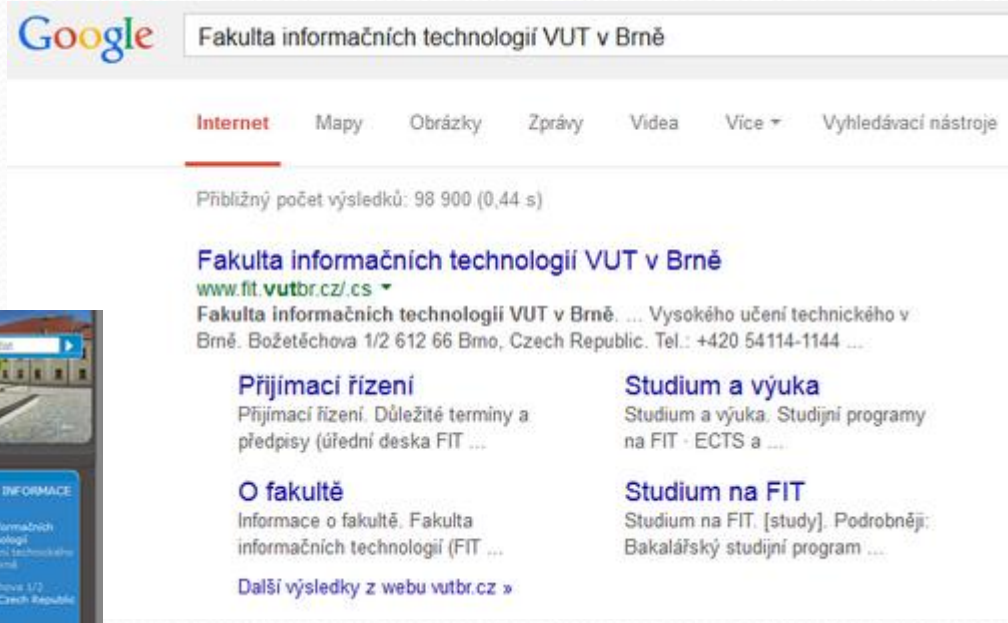
Modern Theoretical Computer Science
2014 / 2015

Author: Ing. Petr Loukota
Advisor: Doc. Ing. Jaroslav Zendulka, CSc.

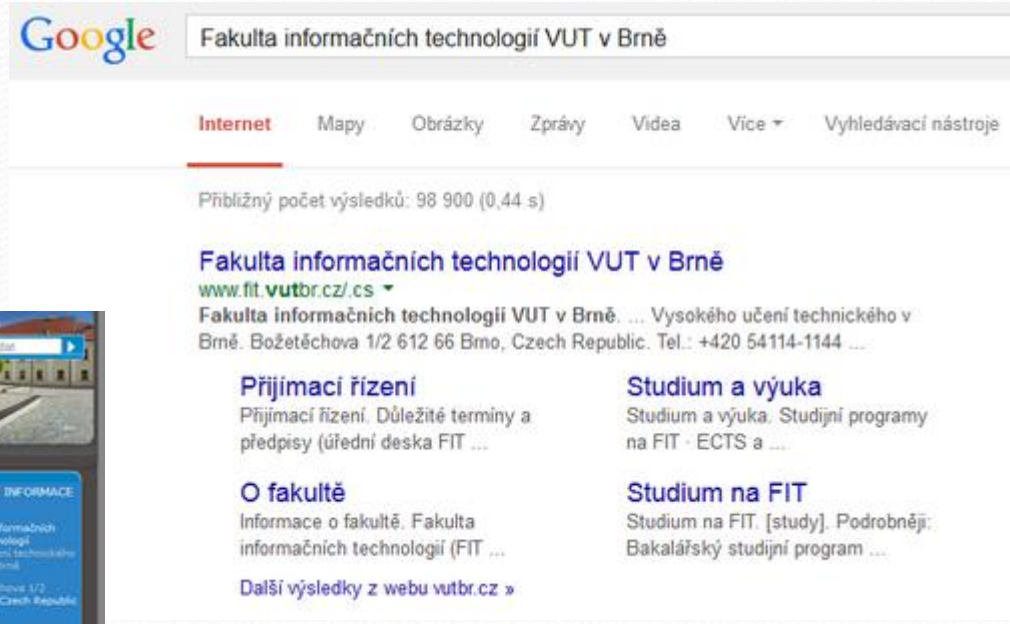
Contents

- Search Engines and Web Page Classification
- Stemming, difference between Stemming and Lemmatization
- Classification of Stemming Algorithms
- Truncating Methods, Statistical Methods, Inflectional and Derivational Methods, Corpus Based Methods and Context Sensitive Methods
- Porter Stemming Algorithm
- Errors in Stemming

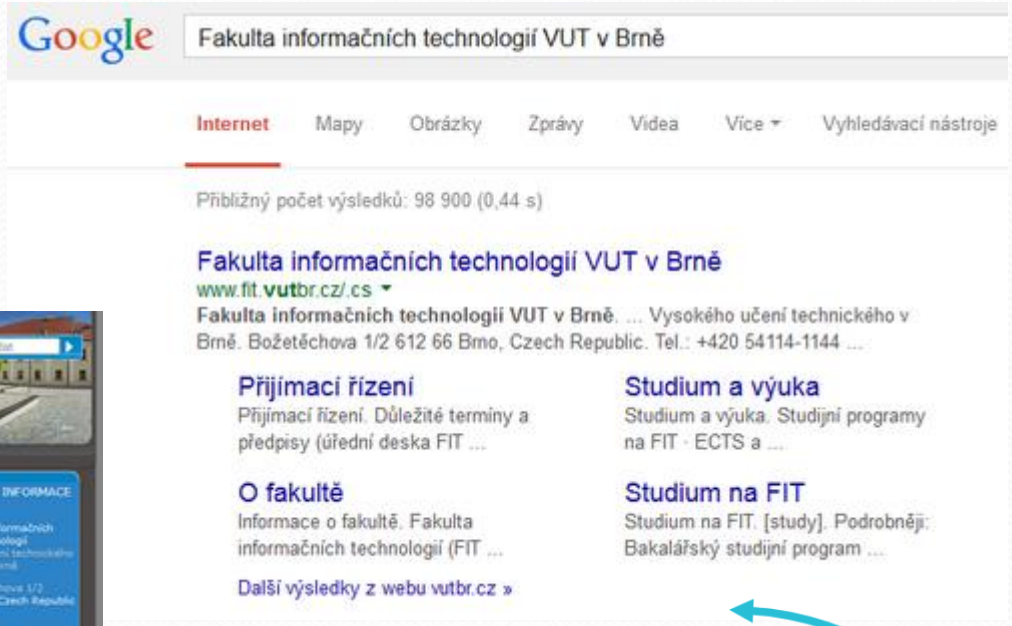
How Search Engines work



How Search Engines work



How Search Engines work



Web Page Classification

- Classification is a process of assigning some of pre-defined class labels to a web document.
- Classification can take both visual and textual information into consideration.
- Two standard pre-processing procedures are applied on the text obtained from a web page – stop words removal and stemming.

Stemming

- Stemming is a pre-processing step in Text Mining applications as well as a very common requirement of Natural Language procession functions.
- It is very important in most of the Information Retrieval systems.
- The main purpose of stemming is to reduce different grammatical forms of a word like its noon, adjective, verb, adverb etc. to its root form.

Stemming

- Stemming is a pre-processing step in Text Mining applications as well as a very common requirement of Natural Language procession functions.
- It is very important in most of the Information Retrieval systems.
- The main purpose of stemming is to reduce different grammatical forms of a word like its noon, adjective, verb, adverb etc. to its root form.

Stemming and Lemmatizing

Stemming:

- In stemming the ‘stem’ is obtaining after applying a set of rules but without bothering about part of speech or the context of the word occurrence.

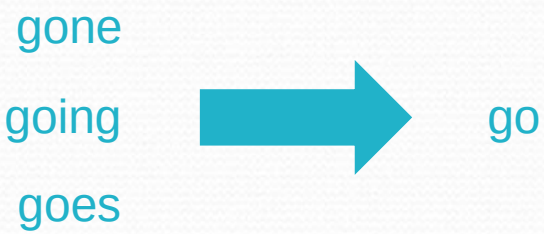
Lemmatization:

- Lemmatization deals with obtaining the ‘lemma’ of a word which involves reducing the word forms to its root form after understanding the part of speech and the context of the word in the given sentence.

Stemming and Lemmatizing

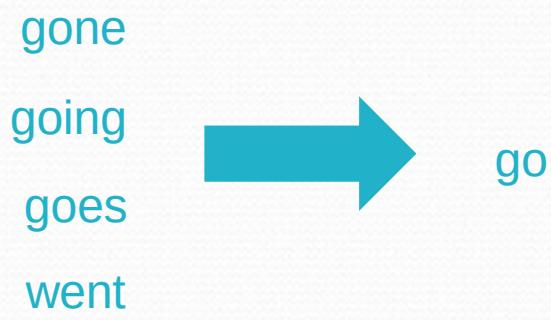
Stemming:

- In stemming the 'stem' is obtaining after applying a set of rules but without bothering about part of speech or the context of the word occurrence.



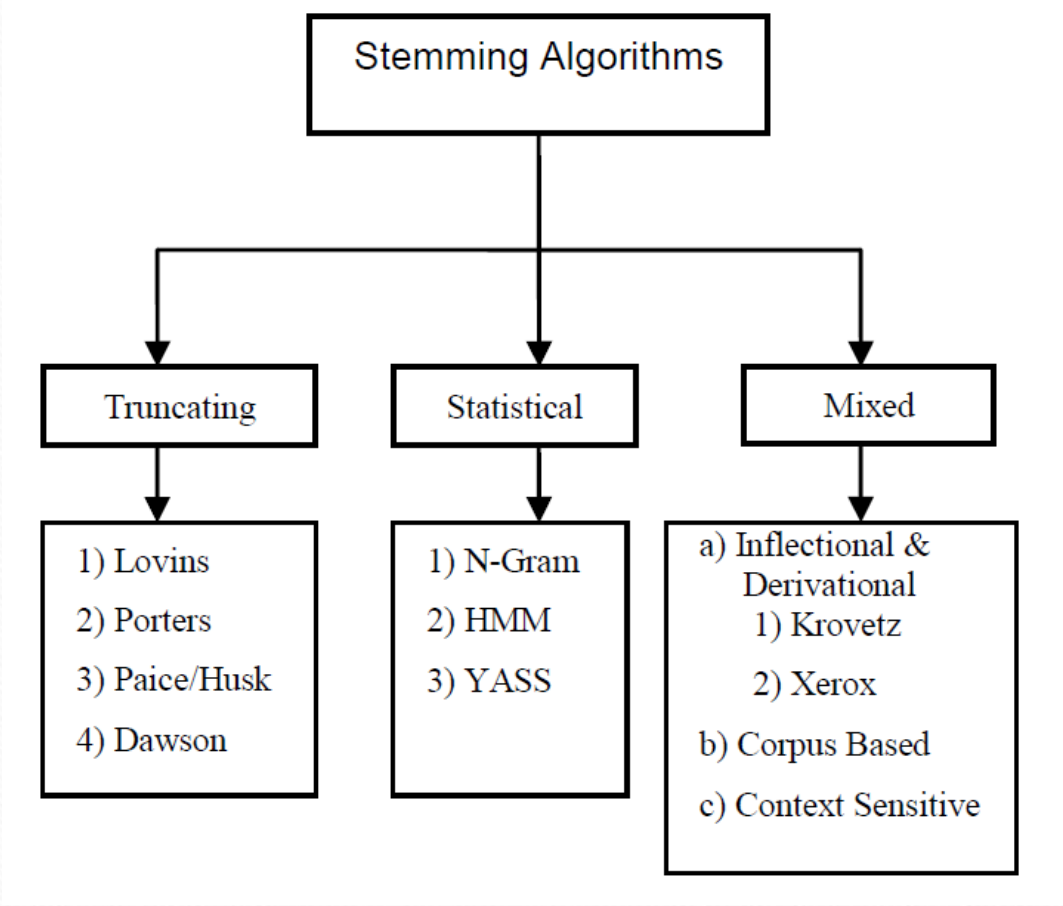
Lemmatization:

- Lemmatization deals with obtaining the 'lemma' of a word which involves reducing the word forms to its root form after understanding the part of speech and the context of the word in the given sentence.



Classification of Stemming Algorithms

- Since the meaning is same but the word form is different it is necessary to identify each word form with its base form.
- To do this a variety of Stemming Algorithms have been developed.



Truncating Methods

- These methods are related to removing the suffixes or prefixes (commonly known as affixes) of a word.

Truncating Methods

- These methods are related to removing the suffixes or prefixes (commonly known as affixes) of a word.

Porter Stemming Algorithm:

- It has five steps, and within each step, rules are applied until one of them passes the condition.
- The rule look like the following: <condition> <suffix> -> <new suffix>

Example: (m > 0) EED -> EE

This rule means „if the word has at least one vowel and consonant plus EED ending, change the ending to EE“.

So „agreed“ becomes „agree“ while „feed“ remains unchanged.

Statistical Methods

- These are the Stemmers who are based on statistical analysis and techniques. Most of the methods remove the affixes but after implementing some statistical procedure.

Statistical Methods

N-Gram Stemmer:

- String-similarity approach is used to convert word inflation to its stem.
- An n-gram is a set of consecutive characters extracted from a word.
- Similar word will have a high proportion of n-grams in common.

Example: the words „statistics“ and „statistical“ for n equals to 2

The diagrams: st, ta, at, ti, is, st, ti, ic, cs

st, ta, at, ti, is, st, ti, ic, ca, al

Advantage: It is language independent.

Disadvantage: It requires a significant amount of memory and storage

Statistical Methods

N-Gram Stemmer:

- String-similarity approach is used to convert word inflation to its stem.
- An n-gram is a set of consecutive characters extracted from a word.
- Similar word will have a high proportion of n-grams in common.

Example: the words „statistics“ and „statistical“ for n equals to 2

The diagrams: st, ta, at, ti, is, st, ti, ic, cs

st, ta, at, ti, is, st, ti, ic, ca, al

Advantage: It is language independent.

Disadvantage: It requires a significant amount of memory and storage

Inflectional and Derivational Methods

- It involves both the inflectional as well as the derivational morphology analysis.
- In case of inflectional the word variants are related to the language specific syntactic variations like plural, gender, case etc. whereas in derivational the world variants are related to the part of speech of a sentence where the word occurs.

Inflectional and Derivational Methods

- It involves both the inflectional as well as the derivational morphology analysis.
- In case of inflectional the word variants are related to the language specific syntactic variations like plural, gender, case etc. whereas in derivational the world variants are related to the part of speech of a sentence where the word occurs.

- **Derivational** affixes do **frequently change the category** of the stem they attach to:

  write (verb) -> writ**er** (noun), break (verb) -> break**able** (adjective)
- **inflectional** affixes **never change the category** of the stems they attach to:

  cow (noun) -> cow**s** (noun), hate (verb) -> hat**ed** (verb)

Corpus Based Methods

- This approach tries to overcome some of the drawback of Porter Stemmer.
  - For example, the words „policy“ and „police“ are conflated though they have a different meaning but the words „index“ and „indices“ are not conflated though they have the same root.
  - Porter Stemmer also generates stems which are not real words like „iteration“ becomes „iter“ and „general“ becomes „gener“.
- Corpus based stemming refers to automatic modification of words that have resulted in a common stem to suit the characteristics of a given text corpus using statistical methods.
- Using this concept some of the over-stemming and under-stemming drawbacks are resolved, e. g. „policy“ and „police“ will no longer be conflated.

Context Sensitive Methods

- This method was proposed by Funchun Peng.
- Unlike the usual method where stemming is done before indexing a document, over here for a Web Search, context sensitive analysis is done using statistical modeling on the query side.
- First, based on statistical language modeling, context sensitive analysis on the query side is performed. This predicts which of its morphological variants is useful to expand a query term with before submitting the query to the search engine.
- Second, context sensitive document matching for those expanded variants is performed.

Porter Stemming Algorithm

- One of the most popular Stemming Algorithms proposed in 1980 by Martin Porter
- Many modifications and enhancements have been done and suggested on the basic algorithm.
- It is based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes.
- It has five steps, and within each step, rules are applied until one of them passes the conditions.
- If a rule is accepted, the suffix is removed accordingly.

Porter Stemming Algorithm

Consonant – a letter other than A, E, I, O, U, and Y preceded by a consonant

Example: TOY – consonants are T and Y, SYZYGY – consonants are S, Z, G

Vowel – a letter that is not a consonant

Consonant will be denoted by c, vowel will be denoted by v.

A list ccc of length greater than 0 will be denoted by C.

A list vvv of length greater than 0 will be denoted by V.

Any word has one of the four forms:

CVCV … C, CVCV … V, VCVC … C, VCVC …. V -> [C]VCVC … [V] -> [C](VC){m}[V]

Porter Stemming Algorithm

The rules for removing a suffix in the form: (condition) S1 -> S2

Example: (m > 1) EMENT -> REPLACEMENT -> REPLAC

The condition may also contain expression with and, or and not and the following:

*S – the stem ends with S (similarly for other letters)

v – the stem contains a vowel

*d – the stem ends with a double consonant (e. g. –TT, -SS)

*o – the stem ends cvc, where the second c is not W, X or Y (e. g. –WIL, -HOP)

Only one rule is obeyed – the one **with the longest matching** S1 for the given word.

Porter Stemming Algorithm – Step 1

a) SSES -> SS
IES -> I
SS -> SS
S ->

caresses -> caress
ponies -> poni
caress -> caress
cats -> cat

b) (m > 0) EED -> EE
(*v*) ED ->
(*v*) ING ->

agreed-> agree
plastered -> plaster
motoring -> motor

If the second or third of the rules in b) is successful, the following is done:

AT -> ATE
BL -> BLE
IZ -> IZE
(*d and not (*L or *S or *Z)) -> single letter
(m = 1 and *o) -> E

conflat(ed) -> conflate
troubl(ed) -> trouble
siz(ed) -> size
hopp(ing) -> hop
fil(ing) -> file

Porter Stemming Algorithm – Step 1

c) (*v*) Y -> I happy -> happi

- Step 1 deals with plurals and past participles. The subsequent steps are much more straightforward.

Porter Stemming Algorithm – Step 2

(m > 0) ATIONAL -> ATE	relational -> relate
(m > 0) TIONAL -> TION	conditional -> condition
(m > 0) ENCI -> ENCE	valenci -> valence
(m > 0) ANCI -> ANCE	hesitanci -> hesitance
(m > 0) IZER -> IZE	digitizer -> digitize
(m > 0) ABLI -> ABLE	conformabli -> conformable
(m > 0) ALLI -> AL	radicalli -> radical
(m > 0) ENTLI -> ENT	differentli -> different
(m > 0) ELI -> E	vileli -> vile
(m > 0) OUSLI -> OUS	analogousli -> analogous
(m > 0) IZATION -> IZE	vietnamization -> vietnamize
(m > 0) ATION -> ATE	predication -> predicate
(m > 0) ATOR -> ATE	operator -> operate
(m > 0) ALISM -> AL	feudalism -> feudal

Porter Stemming Algorithm – Step 2

(m > 0) IVENESS -> IVE	decisiveness -> decisive
(m > 0) FULNESS -> FUL	hopefulness -> hopeful
(m > 0) OUSNESS -> OUS	callousness -> callous
(m > 0) ALITI -> AL	formaliti -> formalal
(m > 0) IVITI -> IVE	sensitiviti -> sensitive
(m > 0) BILITI -> BLE	sensibiliti -> sensible

Porter Stemming Algorithm – Step 3

(m > 0) ICATE -> IC
(m > 0) ATIVE ->
(m > 0) ALIZE -> AL
(m > 0) ICITI -> IC
(m > 0) ICAL -> IC
(m > 0) FUL ->
(m > 0) NESS ->

triplicate -> triplic
formative -> form
formalize -> formal
electriciti -> electric
electrical -> electric
hopeful -> hope
goodness -> good

Porter Stemming Algorithm – Step 4

(m > 1) AL ->	revival -> reviv
(m > 1) ANCE ->	allowance -> allow
(m > 1) ENCE ->	inference -> infer
(m > 1) ER ->	airliner -> airlin
(m > 1) IC ->	gyroscopic -> gyroscop
(m > 1) ABLE ->	adjustable -> adjust
(m > 1) ANT ->	irritant -> irrit
(m > 1) EMENT ->	replacement -> replac
(m > 1) MENT ->	adjustment -> adjust
(m > 1) ENT ->	dependent -> depend
(m > 1 and (*S or *T)) ION ->	adoption -> adopt
(m > 1) OU ->	homologou -> homolog
(m > 1) ISM ->	communism -> commun
(m > 1) ATE ->	activate -> activ

Porter Stemming Algorithm – Step 4

(m > 1) ITI ->	angulariti -> angular
(m > 1) OUS ->	homologous -> homolog
(m > 1) IVE ->	effective -> effect
(m > 1) IZE ->	bowdlerize -> bowdler

- The suffixes are now removed. All that remains is a little tidying up.

Porter Stemming Algorithm – Step 5

a) (m > 1) E ->	probate -> probat
(m = 1 and not *o) E ->	cease -> ceas
b) (m > 1 and *d and *L) -> single letter	controll -> control

- The algorithm is careful not to remove a suffix when the stem is too short, the length of the stem being given by its measure, m. There is no logistic basis for this approach.

Porter Stemming Algorithm

Advantages:

- Its simplicity and speed.
- Produces the best output as compared to other Stemmers.
- Less error rate.
- The Snowball Stemmer Framework designed by Porter is language independent approach to Stemming.

Limitations:

- The produced stems are not always real words.
- It has at least five steps and sixty rules and hence is time consuming.

Errors in Stemming

- There are mainly two errors in Stemming – over-stemming and under-stemming.

Over-stemming:

- Two words with different stems are stemmed to the same root.
- This is also known as a false positive.

Under-stemming:

- Two words that should be stemmed to the same root are not.
- This is also know as a false negative.

References

- **Vladimír Bartík, Radek Burget**
  Two-phase Categorization of Web Documents
  Brno University of Technology
- **Anjali Ganesh Jivani**
  A Comparative Study of Stemming Algorithms
  The Maharaja Sayajirao University of Baroda
- **William B. Frakes**
  Information retrieval: data structures and algorithms

References

- **Fuchun Peng**
  Context Sensitive Stemming for Web Search
  Sunnyvale, California
- **Martin Porter**
  An algorithm for suffix stripping

Thank you for your attention