02. Text Preprocessing

Computational Methods for Text Analysis

Пестова Алена

НИУ ВШЭ Санкт-Петербург

19 сентября 2022 г.

Tokenization: N-grams

N consecutive words.

Unigrams

Восторг внезапный ум пленил .

Bigrams

Восторг внезапный ум пленил <.>

Trigrams

<s> Восторг внезапный ум пленил <.>

Tokenization: N-grams

N consecutive words.

Unigrams

Восторг внезапный ум пленил .

Bigrams

Восторг внезапный ум пленил <.>

Trigrams

<s> Восторг внезапный ум пленил <.>

Tokenization: N-grams

N consecutive words.

Unigrams

Восторг внезапный ум пленил .

Bigrams

Восторг внезапный ум пленил <.>

Trigrams

<s> Восторг внезапный ум пленил <.>

Tokenization: N-grams

N consecutive words.

Unigrams

Восторг внезапный ум пленил .

Bigrams

Восторг внезапный ум пленил <.>

Trigrams

<s> Восторг внезапный ум пленил <.>

Tokenization: N-grams

N consecutive words.

Unigrams

Восторг внезапный ум пленил .

Bigrams

Восторг внезапный ум пленил <.>

Trigrams

<s> Восторг внезапный ум пленил <.>

N-grams

We can make the same frequency lists with N-grams as we did with single words.

How to count words

In order to study word distribution, we need to count the number of occurrences (tokens) of each word (type) in the text.

Qs:

- What is a token? (What should be counted, and what shouldn't?)
- What tokens should be counted as the same *type*?

Stemming

Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.

rule-based algorithms

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Stemming

- Overstemming - nonsensical stems/different words with the same stems (ex.: university, universal, universities, and universe.)
- Understemming - several words that actually are forms of one another (ex.: data, datum -> dat, datu)

Stemming for Russian

- Porter stemmer
- Stemka

Lemmatization

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.

- uses morhological analysis
- resolves a word to its dictionary form (lemma)
- saves the part of speech tag (POS-tag) of the word
- needs more information, more complex algorithms then for stemming

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

Morphological analysis for Russian

- Mystem (lemmas, POS-tags, grammatical forms, works with homonyms), R, Python or even with command line
- udpipe (+ syntax), R and Python
- Stanza (+ syntax), neural algorithm, Python, slower
- DeepPavlov (+ syntax), neural algorithm, Python, slower
- pymorphy/pymorphy2

Ambiguity — homonymy of lingistic signs

Одно слово или разные?

Косил косой косой косой.

коса=S,жен,неод=твор,ед

косая=S,жен,од=(род,ед|дат,ед|твор,ед|пр,ед)

косой=S,муж,од=им,ед

косой=A=(им,ед,полн,муж|род,ед,полн,жен|
дат,ед,полн,жен|вин,ед,полн,муж,неод|твор,ед,полн,жен|
пр,ед,полн,жен)

Ambiguity — homonymy of lingistic signs

Одно слово или разные?

Косил косой косой косой.

косить=V,несов=прош,ед,изъяв,муж,пе

косой=S,муж,од=им,ед

косой=A=твор,ед,полн,жен

коса=S,жен,неод=твор,ед

Homonyms are words which are identical in sound form and/or spelling but different in meaning.
Examples:

Someone left you a **rose**. - The price **rose** significantly last month.

You can't **park** your car here. - Are you heading to the **park** now?

Preprocessing steps

1. Cleaning the data
2. Tokenization (think about what are the tokens, what to remain and what to drop)
3. Working with words' forms (stemming/lemmatization/nothing) (what parts of speech to remain?)
4. Stopwords (to drop or not to drop, how we define stopwords)

Terminology

corpus — collection of texts

token — unit of an analysis in the text (for ex.: word)

wordform — a word in the text with all its "modifications case, tense, etc.

lexeme — a unit of lexical meaning that underlies a set of words that are related through inflection (word in a dictionary, a set of all word forms)

stemming — cutting the word to its word stem, base or root form

lemmatization — getting the initial form of the word

POS-tag - part of speech tag (ex: verb, noun)

Terminology

corpus size - the number of all the words/tokens in the corpus (not unique tokens

vocabulary - a set of all unique words of the corpus

vocabulary size - the number of the unique words/tokens in the corpus