Preprocessing

Чувакин Сергей

R meetup

30 ноября 2019 г.

Outline

- Vectorization
- Data cleaning
- Tokenization et Normalization
- BOW
- TF-IDF
- N-grams

Vectorization

Векторизация - процесс превращения в текста на естесственном языке в вектора.

- *Rule-based*
- Обучение

Pipline

- Чистка данных («tidy data», tags, stop-words)
- Нормализация (stemming, lemmatization)
- N-grams (?)
- Векторизация (simple bow, tf-idf)
- Уменьшение размерности (hash trick)

Data cleaning

- Убрать служебные слова
- stop-words

Tokenization

What is text?

Токенизация - разделение текста на токены.

- Буквы
- Слова
- Фразы
- Параграфы
- etc...

Tokenization

1. Японский, немецкий, адыгейский, аварский языки.
2. Знаки препинания, апострофы, отрицания.

Normalization

Стемминг - удаление суфиксов и окончаний у слов. Процесс основан на правилах.

Лемматизация - превращение слова в словарную форму. Процесс основанных на словарях.

Stemming

Porter, Martin F. 1980.An algorithm for suffix stripping. Program 14 (3): 130-137.

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Stemming

- Lovins stemmer
- Porter stemmer
- Paice stemmer

Stemming

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 vs. Lemmatization

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

BOW

Steps:

1. Make a dictionary.
2. Create zero vectors.
3. Fill with values corresponding the words.
4. Scoring words

TF-IDF

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

TF-IDF

1. **Term Frequency**: is a scoring of the frequency of the word in the current document.
2. **Inverse Document Frequency**: is a scoring of how rare the word is across documents.

N-grams

Example: «It was the best of times»

1. it was
2. was the
3. the best
4. best of
5. of times

Hash trick

Хэшинг - алгоритм представления строк в виде числа.

1. есть разные имплементации
2. можно варьировать длину и размер массива
3. хорошо подходит для уменьшения размера больших корпусов
4. замена словарю