НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied Mathematics and Business Informatics

INTRODUCTION TO NATURAL LANGUAGE PROCESSING (NLP)

Анализ баз данных в публичном управлении

Кольцов С.Н.

Saint Petersburg, 07.09.2018

ТЕКСТОВЫЕ ДАННЫЕ: ЭЛЕМЕНТЫ

Теоретическую основу автоматической обработки текстов составляет компьютерная лингвистика, в основе которой лежат методы машинного обучения, статистического анализа, модели Маркова, логические модели и модификации этих методов с учетом специфики Больших Данных. Однако для того что бы использовать методы машинного обучения тексты должны быть подвергнуты специальной обработке.

Чтобы автоматически анализировать тексты, надо научить компьютер их понимать, т.е. видеть элементы текстов:

- Слова
- N-граммы, словосочетания
- Предложения
- абзацы

*Некоторые термины:

- Корпус / коллекция: выборка текстов
- Уникальные слова (unique words, types): «виды» слов в коллекции
- Вхождения (instances, tokens): все слова, включая повторяющиеся

ТЕКСТОВЫЕ ДАННЫЕ: ПРЕПРОЦЕССИНГ

Токенизация (парсинг) – это процедура разделение входного текста на составные элементы (слова, знаки препинания, числа и так далее).

Очистка (от нетекстовых элементов, знаков препинания, больших букв и нек. др.) – процедура удаления не важных для анализа элементов текста.

Лемматизация – процедура приведения элементов текста к нормальной форме. Процедура основана на словарях.
Стемминг – процедура приведения элементов текста к нормальной форме без учета словаря. Например некоторые алгоритмы просто отрезают концы слов.

POS (part of speech tagging) – процедура частеречной разметки текста. В ходе процедуры каждому элементу сопоставляется метка (Tag) соответствующей части речи.
Удаление стоп-слов – процедура удаления слов, которые не являются фичами.

Приведение в векторную форму (**word-document matrix**) – процедура создания матрицы фич, в которой строка это документ, колонки слова.

Прикладная и компьютерная лингвистика. Коллективная монография. Под редакцией Николаевой И.С., Митрениной О.В., Ландо Т.М., URSS, Москва, 2017.

ВЕКТОРНАЯ МОДЕЛЬ ТЕКСТА

Строки матрицы: документы
Столбцы матрицы : слова, N-граммы и другие элементы текста (например слова)
Содержимое матрицы: частоты
Размерность: количество признаков (слов)

Частоты:
Абсолютные
Tf-idf (отношение частоты слова в тексте к числу текстов, в котором встречается данное слово)

Термины/ документы	Doc1	Doc2	Doc3
в	5	2	10
время	5	2	0
мост	0	7	8
петербург	5	15	25
разводка	1	4	0

	A	B	C	D	E	F	G
1	myclass	гламур	журнал	фото	стилист	макияж	модель
2	политика	0	0	0	0	0	0
3	политика	0	0	0	0	0	0
4	неопределенный	0	0	0	0	0	0
5	политика	0	0	0	0	0	0
6	мода	0	0	0	0	0	0
7	политика	0	2	0	0	0	1
8	политика	0	0	0	0	0	0
9	политика	0	0	0	0	0	0
10	насилие	0	0	0	0	0	0
11	насилие	0	0	0	0	0	0
12	насилие	0	0	0	0	0	0
13	политика	0	2	0	0	0	0
14	политика	0	0	0	0	0	0
15	насилие	0	0	0	0	0	0
16	политика	0	0	0	0	0	35
17	насилие	0	0	0	0	0	0
18	политика	0	0	1	0	0	0
19	насилие	0	0	0	0	0	0
20	политика	0	0	0	0	0	1
21	история	0	0	0	0	0	0
22	насилие	0	0	0	0	0	0

TF-IDF

TF-IDF (от англ. **TF — term frequency**, **IDF — inverse document frequency**) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

TF (term frequency — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова t_i в пределах отдельного документа.

$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k}$$

где $\mathbf{n_t}$ есть число вхождений слова **t** в документ, а в знаменателе — общее число слов в данном документе.

IDF (*inverse document frequency* — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|}$$

|D| — число документов в коллекции;
— число документов из коллекции D, в которых встречается t

$$\text{tf-idf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

РАСПРЕДЕЛЕНИЕ СЛОВ В ТЕКСТАХ И В ЯЗЫКЕ

$$f(x) \sim Cx^{-\alpha}.$$

- Закон Ципфа
- Распределение Парето
- Степенное распределение
- Все они похожи; о различиях популярно здесь:

Lada Adamic. Zipf, Power-laws and Pareto – a ranking tutorial:
http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html

M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, 2006

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru