Text mining

- Другие названия для Text Mining:
  - Автоматическая обработка текстов
  - Natural Language Processing
  - Компьютерная лингвистика (Computational Linguistics)
- Компьютерная лингвистика (КЛ) — междисциплинарная область, которая возникла на стыке таких наук, как лингвистика, математика, информатика (Computer Science), искусственный интеллект (Artificial Intelligence).
- В своем развитии она до сих пор вбирает и применяет (при необходимости адаптируя) разработанные в этих науках методы и инструменты.

Задачи в области ТМ

- Машинный перевод
- Информационный поиск (Information Retrieval)
- Реферирование текста (Summarization)
  - Аннотирование, ключевые слова
- Задача рубрицирования текста (Text Classification)
- Выделение мнений (Opinion Mining) и анализ тональности текстов (Sentiment Analysis)
- Автоматическая генерация текстов

Специфика анализа текстов

- Вариативность языка (использование близких по смыслу слов, отклонения от нормы, и т.п.)
- Многозначность слов и важность контекста для понимания отдельных слов
- Анализ текстов как последовательностей

Предообработка текстов

1. Удаление символов
2. Удаление стоп-слов
3. Приведение слов к одному регистру
4. Стемминг и лемматизация (приведение к нормальной форме)
5. **Векторизация слов (числовое представление слов)**

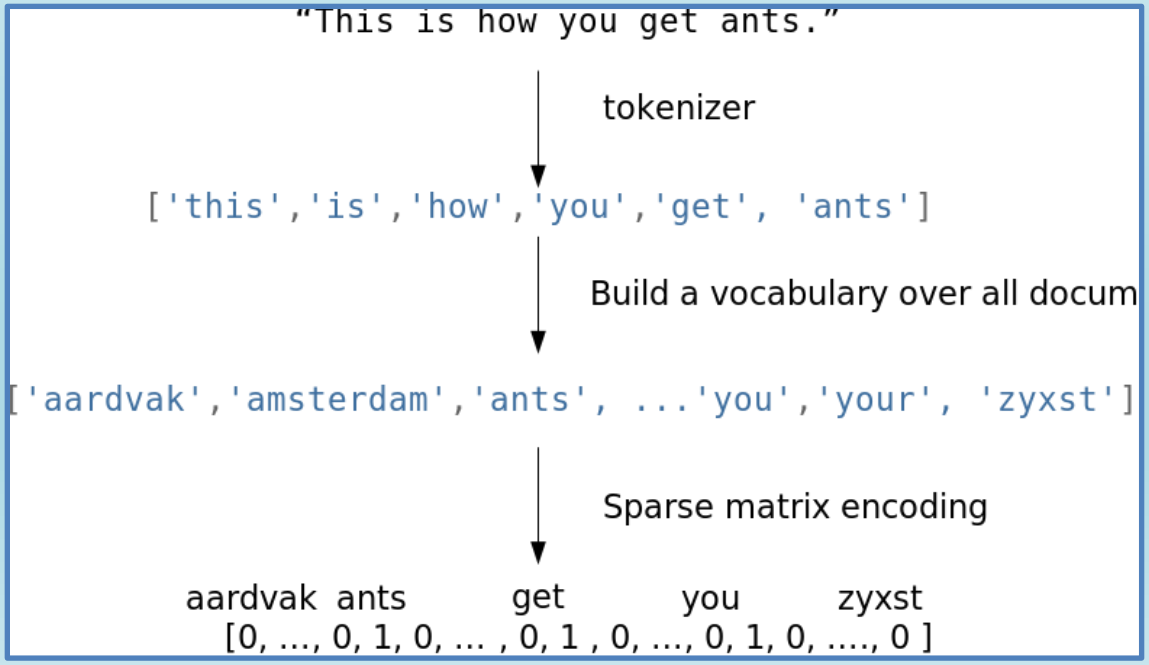
Текст -> Число

- Операция *векторизации*

- Частотная векторизация:
  - One-hot кодирование
  - Bag of word
  - TF\IDF
  - Модификации TF\IDF
- Векторизация с помощью алгоритмов глубокого обучения:
  - Предобученные модели
  - Получение векторов слов на обучающей выборке с помощью нейросетевых алгоритмов

one-shot кодирование

- Каждое слово представляется разреженным булевским вектором
- Размер вектор равен размеру «словаря»
- Вектор содержит единицу только в позиции, соответствующей слову и нули в остальных позициях.
- Такие вектора не позволяют учитывать семантическую близость слов



От one-hot векторов к Bag of Words

- One-hot кодирование

  каждое слово – отдельный вектор

	22	driving	he	old	started	was	when	years
he	0	0	1	0	0	0	0	0
started	0	0	0	0	1	0	0	0
driving	0	1	0	0	0	0	0	0
when	0	0	0	0	0	0	1	0
he	0	0	1	0	0	0	0	0
was	0	0	0	0	0	1	0	0
22	1	0	0	0	0	0	0	0
years	0	0	0	0	0	0	0	1
old	0	0	0	1	0	0	0	0

- Bag of Words

  объединение one-hot векторов

	Sam	started	driving	...	preferred	outdoor	activities
Sent1	1	1	1	...	0	0	0
Sent2	0	0	0	...	0	0	0
Sent3	0	0	0	...	1	1	1

N-grams

- N-граммы слов — это группы по *N* последовательных слов, которые можно извлечь из предложения. Та же идея применима к символам.

«The cat sat on the mat»

Биграммы:

{"The", "The cat", "cat", "cat sat", "sat", "sat on", "on", "on the", "the", "the mat", "mat"}

Триграммы:

{"The", "The cat", "cat", "cat sat", "The cat sat",
"sat", "sat on", "on", "cat sat on", "on the", "the",
"sat on the", "the mat", "mat", "on the mat"}

Векторизация TF-IDF

1) Вычисляется частота термина TF (term frequency) – оценка важности слова t в пределах одного документа d.

$$TF = \frac{C_{t,d}}{C_d}$$

где $C_{t,d}$ – сколько раз слово t встречается в документе d;

C_d – общее число слов в документе.

2) Вычисляется обратная частота документа IDF (inverse document frequency) – инверсия частоты, с которой слово t встречается в документах коллекции. IDF уменьшает вес общеупотребительных слов.

$$ITF = log\ \frac{|D|}{D_t}$$

где $|D|$ - общее количество документов в коллекции;

D_t – количество всех документов, в которых встречается слово t.

Векторизация TF-IDF

3) Итоговый вес слова t в документе d относительно всей коллекции документов вычисляется по формуле:

$$V_{t,d} = TF * IDF$$

Таким образом, большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употребления в других документах.

Модифицированная оценка Tf-Idf

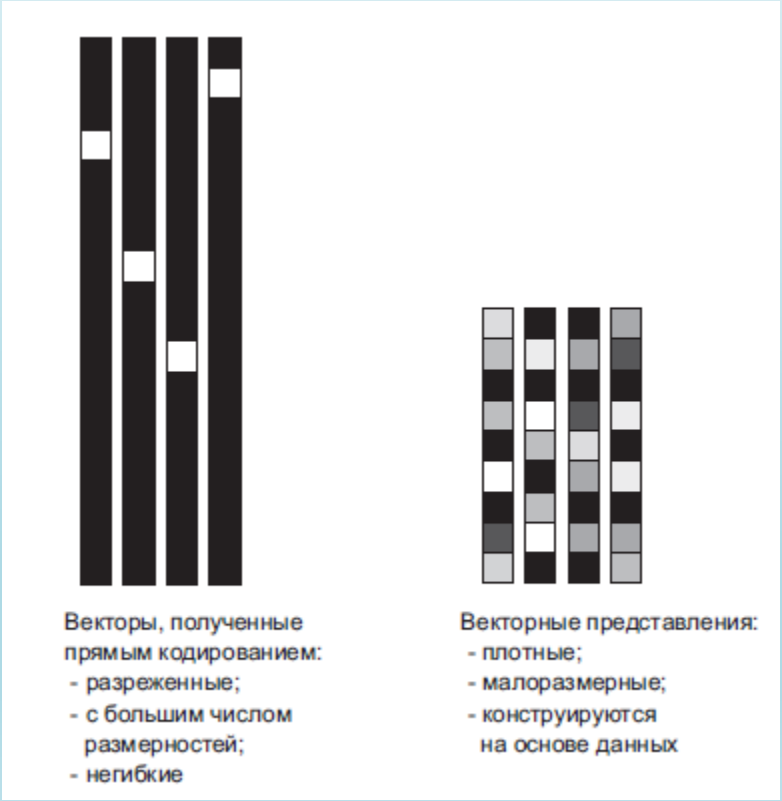
- В задаче классификации текстов также используются модифицированные оценки Tf\Idf, которые учитывают специфичность «терминов» не для всего корпуса документов, а для отдельной категории.
- Например, для случая двух классов («положительные» и «отрицательные» тексты) можно использовать модифицированную оценку:

$$V_{t,d} = C_{t,d} * log\left(\frac{|N| * P_t}{|P| * N_t}\right)$$

- где $C_{t,d}$ – количество раз слово t встречается в документе d;
- $|P|$ - количество документов положительной тональности;
- $|N|$ - количество документов отрицательной тональности;
- P_t – количество документов положительной тональности, в которых встречается слово t;
- N_t – количество документов отрицательной тональности, в которых встречается слово t.

Word Embeddings

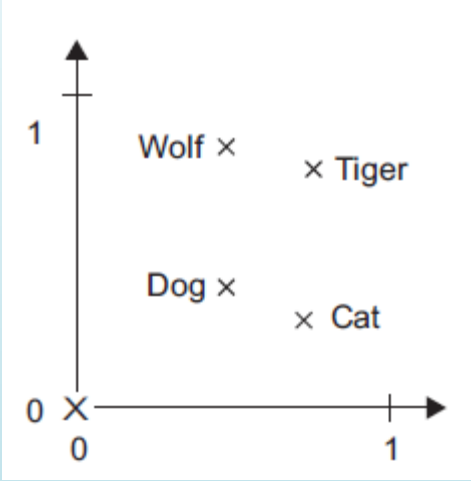
- Векторное представление (word embeddings) – числовые вектора, соответствующие словам

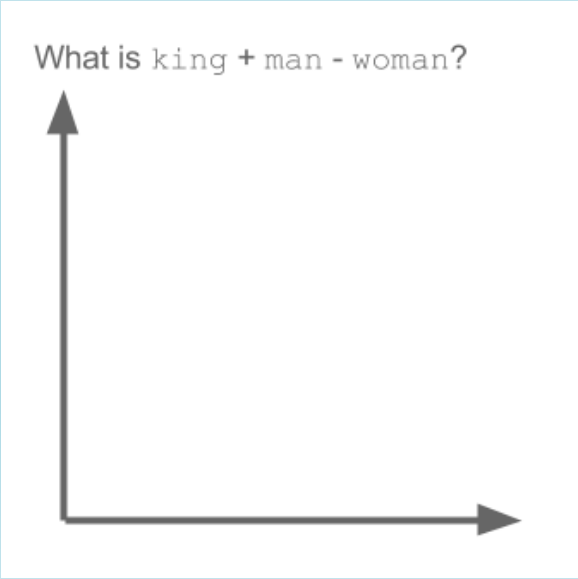


Получить векторные представления слов можно двумя способами:
1. Конструировать векторные представления в процессе решения основной задачи (как правило, при обучении нейронной сети).
2. Загрузить в модель векторные представления, полученные с использованием другой задачи машинного обучения (*предварительно обученные векторные представления слов)*.

Векторные представления

- Считается, что геометрические отношения между числовыми векторами слов должны отражать семантические связи между соответствующими им словами
  - Чем меньше расстояние между векторами, тем ближе слова по смыслу
  - Направления в пространстве векторов также связаны с семантическими связями





Предобученные векторные представления

- Широко используются предварительно сформированные векторные представления, хорошо организованные и обладающие полезными свойствами, которые охватывают основные аспекты языковой структуры.
- В применении предобученных векторных представлений есть смысл при отсутствии достаточного объема данных для выделения хороших признаков.
- Предобученные векторные представления построены:
  - с использованием статистики встречаемости слов;
  - с применением нейронных сетей;
- Распространенные модели:
  - модель word2Vec (2013, Tomas Mikolov, Google);
  - модель Glove (2014, Stanford)
  - модель fasttext