

Обработка естественного языка

Владимир Опанасенко

NewProLab

Москва
апрель 2021 г.

① Что такое компьютерная лингвистика?

- ① Что такое компьютерная лингвистика?
- ② Лингвистические дисциплины

- ① Что такое компьютерная лингвистика?
- ② Лингвистические дисциплины
- ③ Задачи компьютерной лингвистики

- ① Что такое компьютерная лингвистика?
- ② Лингвистические дисциплины
- ③ Задачи компьютерной лингвистики
- ④ Лингвистический анализ

- 1 Что такое компьютерная лингвистика?
- 2 Лингвистические дисциплины
- 3 Задачи компьютерной лингвистики
- 4 Лингвистический анализ
- 5 Инструменты лингвистического анализа

- 1 Что такое компьютерная лингвистика?
- 2 Лингвистические дисциплины
- 3 Задачи компьютерной лингвистики
- 4 Лингвистический анализ
- 5 Инструменты лингвистического анализа
- 6 Векторное представление текстов

Что такое компьютерная лингвистика?

- Наша Таня громко плачет

Что такое компьютерная лингвистика?

- Наша Таня громко плачет



Что такое компьютерная лингвистика?

- Наша Таня громко плачет
- Хочу взять ипотеку



Что такое компьютерная лингвистика?

- Наша Таня громко плачет
- Хочу взять ипотеку
- Как оформить карту?



Что такое компьютерная лингвистика?

- Наша Таня громко плачет
- Хочу взять ипотеку
- Как оформить карту?
- Что такое осень?

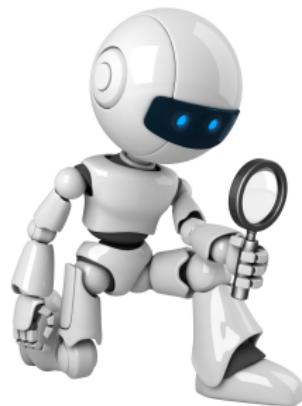


Что такое компьютерная лингвистика?

Как заставить машину понимать язык?

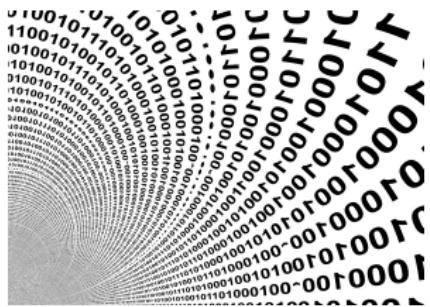
Что такое компьютерная лингвистика?

Как заставить машину понимать язык?

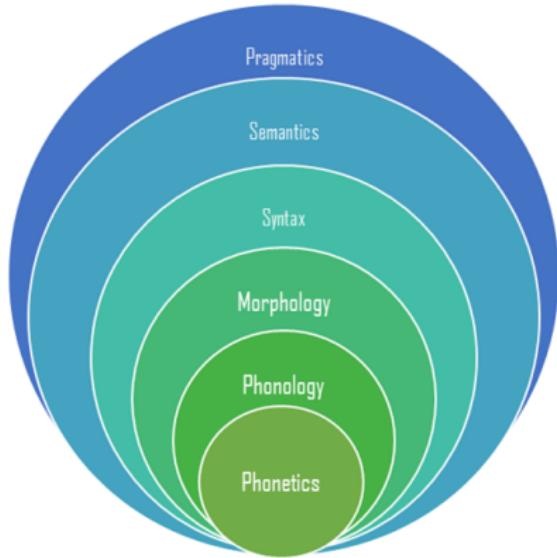


Что такое компьютерная лингвистика?

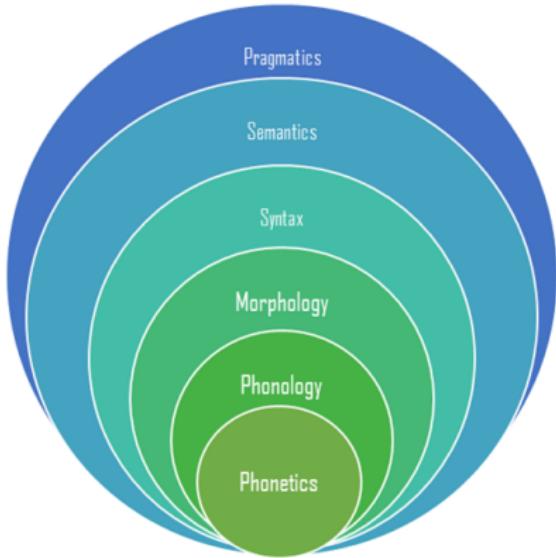
Как заставить машину понимать язык?



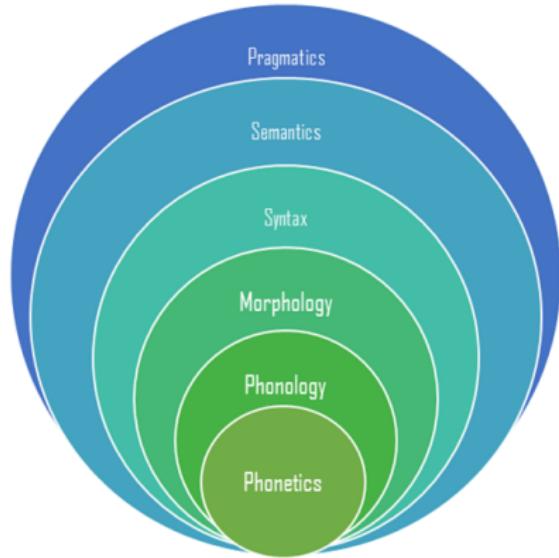
① Фонетика (звуки)



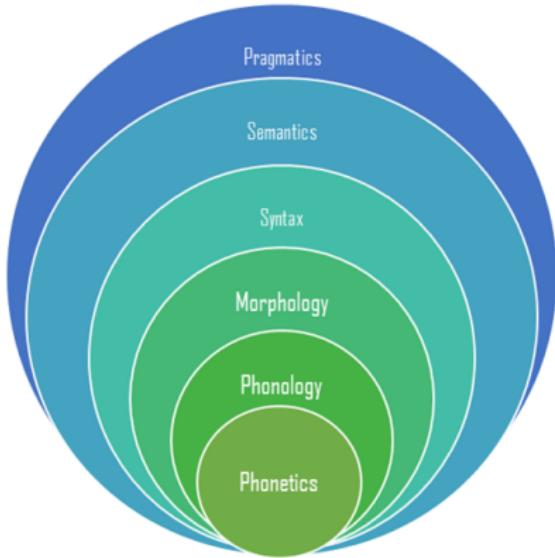
- ① Фонетика
(звуки)
- ② Фонология
(фонемы)



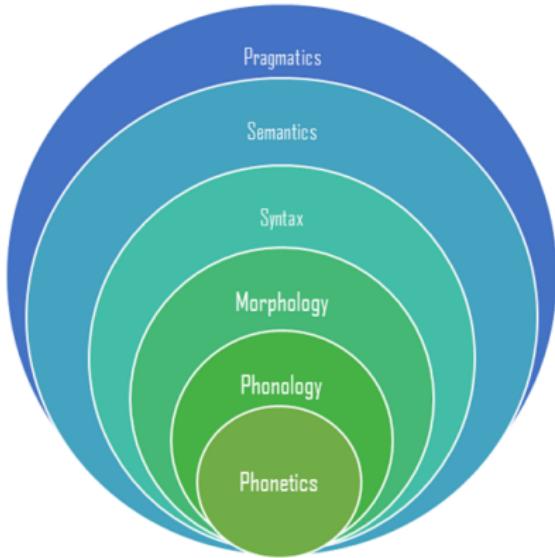
- ① Фонетика
(звуки)
- ② Фонология
(фонемы)
- ③ Морфология
(слова)



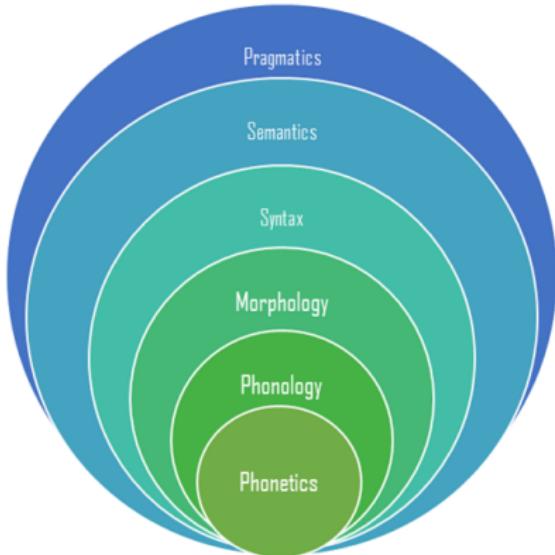
- ① Фонетика
(звуки)
- ② Фонология
(фонемы)
- ③ Морфология
(слова)
- ④ Синтаксис
(предложения)



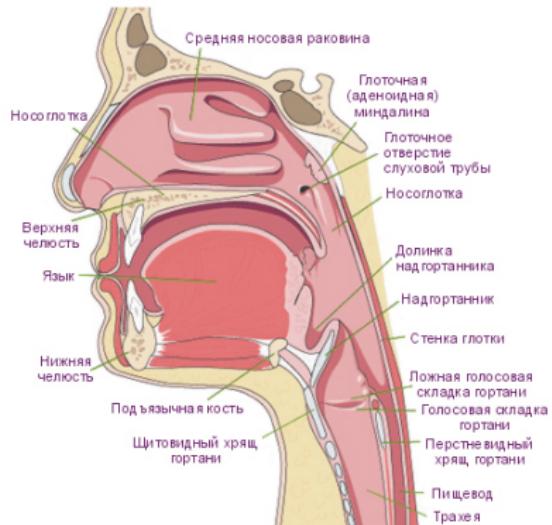
- ① Фонетика
(звуки)
- ② Фонология
(фонемы)
- ③ Морфология
(слова)
- ④ Синтаксис
(предложения)
- ⑤ Семантика
(смысл
предложений)



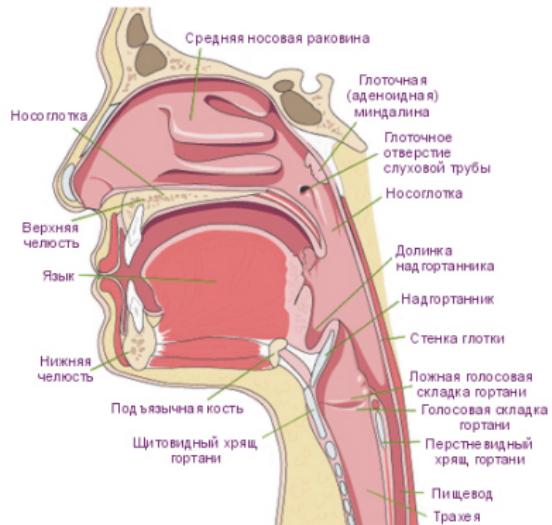
- ① Фонетика
(звуки)
- ② Фонология
(фонемы)
- ③ Морфология
(слова)
- ④ Синтаксис
(предложения)
- ⑤ Семантика
(смысл
предложений)
- ⑥ Прагматика
(смысл текстов)



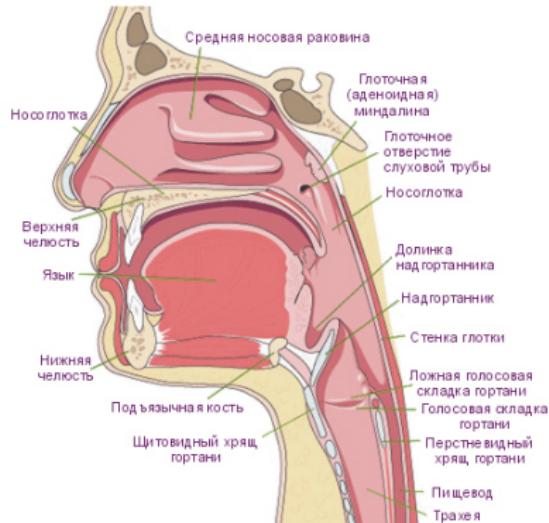
● Звуки речи



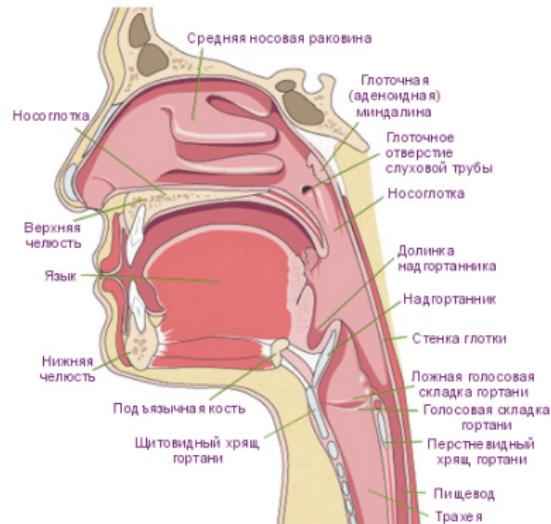
- Звуки речи
- Физика процесса



- Звуки речи
- Физика процесса
 - Как они создаются



- Звуки речи
- Физика процесса
 - Как они создаются
 - Как они воспринимаются



- Звуки речи



- Звуки речи
- Как они собираются в слова



- Звуки речи
- Как они собираются в слова
 - Фонемы и слоги



- Звуки речи
- Как они собираются в слова
 - Фонемы и слоги
 - Транскрипция (звуки - не буквы)



Лингвистика

Морфология

- Внутренняя структура слов



Лингвистика

Морфология

- Внутренняя структура слов
 - Словообразование



Лингвистика

Морфология

- Внутренняя структура слов
 - Словообразование
 - Словоизменение



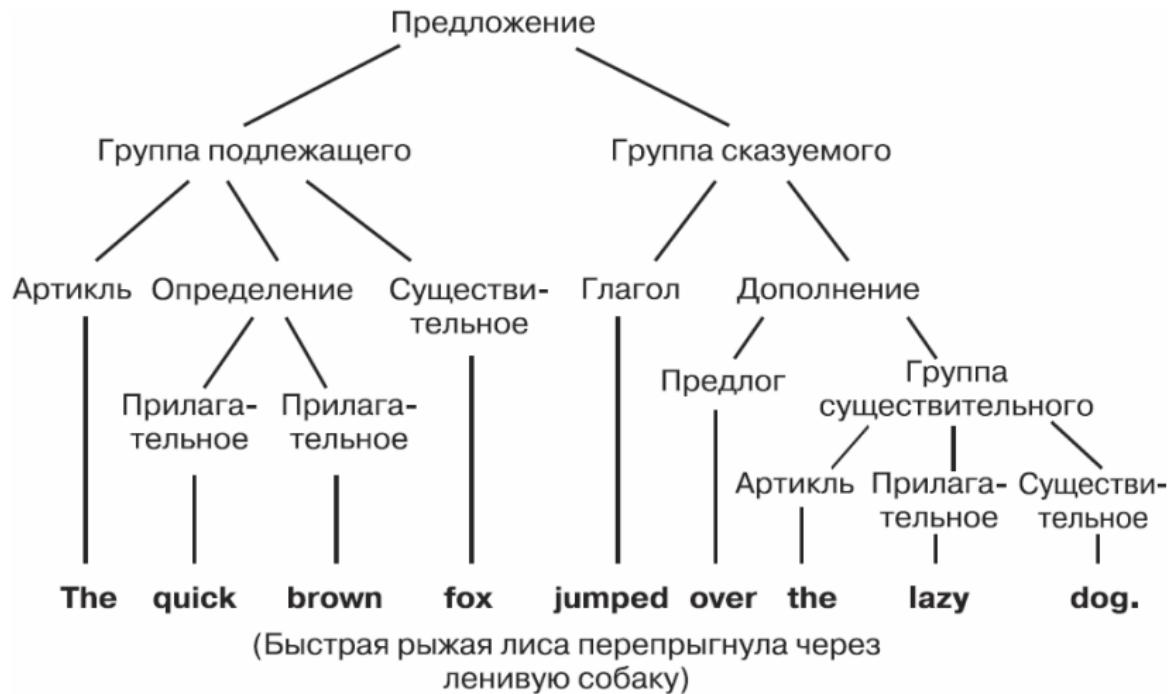
Лингвистика

Морфология

- Внутренняя структура слов
 - Словообразование
 - Словоизменение
- Морфемы



Структура предложений, связи и зависимости между словами



Лингвистика

Семантика

- Значение, смысл
отдельных слов и
предложений

Омонимы
коса

- Значение, смысл отдельных слов и предложений
- Что такое «коса»?



Лингвистика

Прагматика

Sherlock saw the man using binoculars [with binoculars]



see¹_{using binoculars}



see²_{using binoculars}

Лингвистика

Прагматика

- Для понимания высказывания часто недостаточно самого высказывания

Лингвистика

Прагматика

- Для понимания высказывания часто недостаточно самого высказывания
- Нужно понимание контекста, в котором находится автор

Лингвистика

Прагматика

- Для понимания высказывания часто недостаточно самого высказывания
- Нужно понимание контекста, в котором находится автор
 - Ирония: «Спасибо за заботу, ждали всего-то два часа»

Лингвистика

Прагматика

- Для понимания высказывания часто недостаточно самого высказывания
- Нужно понимание контекста, в котором находится автор
 - **Ирония:** «Спасибо за заботу, ждали всего-то два часа»
 - **Неявная просьба:** «Не могли бы вы передать салат?»

Лингвистика

Прагматика

- Для понимания высказывания часто недостаточно самого высказывания
- Нужно понимание контекста, в котором находится автор
 - **Ирония:** «Спасибо за заботу, ждали всего-то два часа»
 - **Неявная просьба:** «Не могли бы вы передать салат?»
 - **Культурные различия:** Froid de canard - утиный холод

Лингвистика

Прагматика

- Для понимания высказывания часто недостаточно самого высказывания
- Нужно понимание контекста, в котором находится автор
 - **Ирония:** «Спасибо за заботу, ждали всего-то два часа»
 - **Неявная просьба:** «Не могли бы вы передать салат?»
 - **Культурные различия:** Froid de canard - утиный холод
 - **Общие предпосылки:** «Эти типы стали есть в прокатном цехе»

- Омонимия

Трудности

- Омонимия
- Анафора

- **Омонимия** – это случайное совпадение слов

- **Омонимия** – это случайное совпадение слов
- Ключ, коса, родник, течь, наряд...

- **Омонимия** – это случайное совпадение слов
- Ключ, коса, родник, течь, наряд...
- **Полисемия** – это наличие у слов разных исторически связанных и близких по смыслу значений



Анафора

Примеры

Анафора – зависимость интерпретации некоторого выражения от другого выражения (антецедента), обычно ранее встречавшегося в тексте

- Коля думает, что он умен

Анафора

Примеры

Анафора – зависимость интерпретации некоторого выражения от другого выражения (антецедента), обычно ранее встречавшегося в тексте

- Коля думает, что он умен
- Если у фермера есть осёл, то он богат

Анафора

Примеры

Анафора – зависимость интерпретации некоторого выражения от другого выражения (антецедента), обычно ранее встречавшегося в тексте

- Коля думает, что он умен
- Если у фермера есть осёл, то он богат
- Я напечатал на принтере документ, а потом разорвал его

Анафора

Примеры

Анафора – зависимость интерпретации некоторого выражения от другого выражения (антецедента), обычно ранее встречавшегося в тексте

- Коля думает, что он умен
- Если у фермера есть осёл, то он богат
- Я напечатал на принтере документ, а потом разорвал его
- *Возможно явное указание в тексте*

- **Согласование слов:** Разработчики презентовали беспилотный автомобиль. Они разрабатывали его в течение трёх лет

- **Согласование слов:** Разработчики презентовали беспилотный автомобиль. Они разрабатывали его в течение трёх лет
- **Синтаксические ограничения:** Брат подарил ему телефон (Брат ≠ ему)

- **Согласование слов:** Разработчики презентовали беспилотный автомобиль. Они разрабатывали его в течение трёх лет
- **Синтаксические ограничения:** Брат подарил ему телефон (Брат \neq ему)
- **Гипотеза центрирования** - если долго говорим о чём-то одном, то местоимение тоже к этому относится

- **Согласование слов:** Разработчики презентовали беспилотный автомобиль. Они разрабатывали его в течение трёх лет
- **Синтаксические ограничения:** Брат подарил ему телефон (Брат \neq ему)
- **Гипотеза центрирования** - если долго говорим о чём-то одном, то местоимение тоже к этому относится
- **Семантические ограничения** - Я напечатал на принтере документ, а потом разорвал его

- В первом приближении - теория порождающих грамматик

- В первом приближении - теория порождающих грамматик
- Но не только

- В первом приближении - теория порождающих грамматик
- Но не только
- Дисциплина прикладной лингвистики, предметом которой является разработка и изучение понятий, образующих основу формального аппарата для описания строения естественных языков

Задачи компьютерной лингвистики

- Машинный перевод



Задачи компьютерной лингвистики

- Машинный перевод
- Полнотекстовый поиск



Задачи компьютерной лингвистики

- Машинный перевод
- Полнотекстовый поиск
- Кластеризация и классификация текстов



Задачи компьютерной лингвистики

- Машинный перевод
- Полнотекстовый поиск
- Кластеризация и классификация текстов
- Оценка тональности текста



Задачи компьютерной лингвистики

- Машинный перевод
- Полнотекстовый поиск
- Кластеризация и классификация текстов
- Оценка тональности текста
- Синтез и распознавание речи



Задачи компьютерной лингвистики

- Машинный перевод
- Полнотекстовый поиск
- Кластеризация и классификация текстов
- Оценка тональности текста
- Синтез и распознавание речи
- Диалоговые системы



Задачи компьютерной лингвистики

- Машинный перевод
- Полнотекстовый поиск
- Кластеризация и классификация текстов
- Оценка тональности текста
- Синтез и распознавание речи
- Диалоговые системы
- Аннотирование, реферирование



Задачи компьютерной лингвистики

- Машинный перевод
- Полнотекстовый поиск
- Кластеризация и классификация текстов
- Оценка тональности текста
- Синтез и распознавание речи
- Диалоговые системы
- Аннотирование, реферирование
- Извлечение фактов, поиск именованных сущностей



Поиск именованных сущностей NER

- Ищем упоминания заранее определенных категорий сущностей в тексте



Поиск именованных сущностей NER

- Ищем упоминания заранее определенных категорий сущностей в тексте
 - Имена, даты, адреса...



Поиск именованных сущностей NER

- Ищем упоминания заранее определенных категорий сущностей в тексте
 - Имена, даты, адреса...
 - Правила и регулярные выражения



Поиск именованных сущностей NER

- Ищем упоминания заранее определенных категорий сущностей в тексте
- Имена, даты, адреса...
- Правила и регулярные выражения
 - Томита-парсер



Поиск именованных сущностей

NER

- Ищем упоминания заранее определенных категорий сущностей в тексте
- Имена, даты, адреса...
- Правила и регулярные выражения
 - Томита-парсер
 - Natasha



Поиск именованных сущностей

NER

- Ищем упоминания заранее определенных категорий сущностей в тексте
- Имена, даты, адреса...
- Правила и регулярные выражения
 - Томита-парсер
 - Natasha
- Разметка и машинное обучение



Поиск именованных сущностей

NER

- Ищем упоминания заранее определенных категорий сущностей в тексте
- Имена, даты, адреса...
- Правила и регулярные выражения
 - Томита-парсер
 - Natasha
- Разметка и машинное обучение
 - DeepPavlov



- Немного истории

- Немного истории
 - Шифр Цезаря и криптоанализ (IX в.)

- Немного истории
 - Шифр Цезаря и криптоанализ (IX в.)
 - Анализ древних языков

- Немного истории
 - Шифр Цезаря и криптоанализ (IX в.)
 - Анализ древних языков
- Важно наличие большого количества текстов на анализируемом языке

- Немного истории
 - Шифр Цезаря и криптоанализ (IX в.)
 - Анализ древних языков
- Важно наличие большого количества текстов на анализируемом языке
- Корпусная лингвистика

Корпусная лингвистика

- Корпус - подобранный и обработанный по определённым правилам совокупность текстов



Корпусная лингвистика

- Корпус - подобранный и обработанный по определённым правилам совокупность текстов
 - Лингвистический корпус - корпус, в котором тексты снабжены лингвистической разметкой



Корпусная лингвистика

- Корпус - подобранный и обработанный по определённым правилам совокупность текстов
 - Лингвистический корпус - корпус, в котором тексты снабжены лингвистической разметкой
 - Проблемы



Корпусная лингвистика

- Корпус - подобранная и обработанная по определённым правилам совокупность текстов
 - Лингвистический корпус - корпус, в котором тексты снабжены лингвистической разметкой
 - Проблемы
 - Представительность



Корпусная лингвистика

- Корпус - подобранная и обработанная по определённым правилам совокупность текстов
 - Лингвистический корпус - корпус, в котором тексты снабжены лингвистической разметкой
 - Проблемы
 - Представительность
 - Разметка



Корпусная лингвистика

- Корпус - подобранный и обработанный по определенным правилам совокупность текстов
 - Лингвистический корпус - корпус, в котором тексты снабжены лингвистической разметкой
 - Проблемы
 - Представительность
 - Разметка
 - Балансировка



Корпусная лингвистика

- Корпус - подобранный и обработанный по определённым правилам совокупность текстов
 - Лингвистический корпус - корпус, в котором тексты снабжены лингвистической разметкой
 - Проблемы
 - Представительность
 - Разметка
 - Балансировка
 - Авторские права



- Многоязычные



- Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)



Корпусная лингвистика

- Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)
 - Уппсальский корпус (1980-е, ≈ 1 млн. слов)



- Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)
 - Уппсальский корпус (1980-е, ≈ 1 млн. слов)
 - British National Corpus (1994, ≈ 100 млн. слов)



• Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)
 - Уппсальский корпус (1980-е, ≈ 1 млн. слов)
 - British National Corpus (1994, ≈ 100 млн. слов)
 - НКРЯ (ruscorpora.ru, более 600 млн. слов)



- Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)
 - Уппсальский корпус (1980-е, ≈ 1 млн. слов)
 - British National Corpus (1994, ≈ 100 млн. слов)
 - НКРЯ (ruscorpora.ru, более 600 млн. слов)
 - OpenCorpora (opencorpora.org, ≈ 1 млн. слов), полностью размечен



Корпусная лингвистика

- Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)
 - Уппсальский корпус (1980-е, ≈ 1 млн. слов)
 - British National Corpus (1994, ≈ 100 млн. слов)
 - НКРЯ (ruscorpora.ru, более 600 млн. слов)
 - OpenCorpora (opencorpora.org, ≈ 1 млн. слов), полностью размечен

- Параллельные



- Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)
 - Уппсальский корпус (1980-е, ≈ 1 млн. слов)
 - British National Corpus (1994, ≈ 100 млн. слов)
 - НКРЯ (ruscorpora.ru, более 600 млн. слов)
 - OpenCorpora (opencorpora.org, ≈ 1 млн. слов), полностью размечен

- Параллельные

- УМС (\approx 100 тыс. предложений)



- Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)
 - Уппсальский корпус (1980-е, ≈ 1 млн. слов)
 - British National Corpus (1994, ≈ 100 млн. слов)
 - НКРЯ (ruscorpora.ru, более 600 млн. слов)
 - OpenCorpora (opencorpora.org, ≈ 1 млн. слов), полностью размечен

- Параллельные

- UMC (\approx 100 тыс. предложений)
 - Яндекс.Перевод (1 млн. пар предложений)



Корпусная лингвистика

- Моноязычные

- Brown corpus (1967, ≈ 1 млн. слов)
 - Уппсальский корпус (1980-е, ≈ 1 млн. слов)
 - British National Corpus (1994, ≈ 100 млн. слов)
 - НКРЯ (ruscorpora.ru, более 600 млн. слов)
 - OpenCorpora (opencorpora.org, ≈ 1 млн. слов), полностью размечен

- Параллельные

- UMC (\approx 100 тыс. предложений)
 - Яндекс.Перевод (1 млн. пар предложений)
 - Europarl



- Частота слова f обратно пропорциональна его рангу r в отсортированном по частотности списке слов

$$f \sim \frac{1}{r}$$

- Частота слова f обратно пропорциональна его рангу r в отсортированном по частотности списке слов

$$f \sim \frac{1}{r}$$

- Иначе, существует константа K такая, что

Закон Ципфа

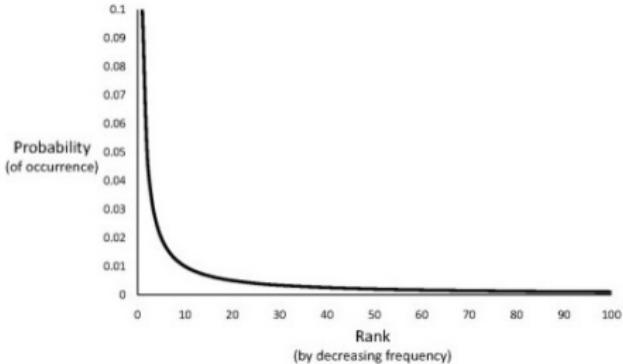
- Частота слова f обратно пропорциональна его рангу r в отсортированном по частотности списке слов

$$f \sim \frac{1}{r}$$

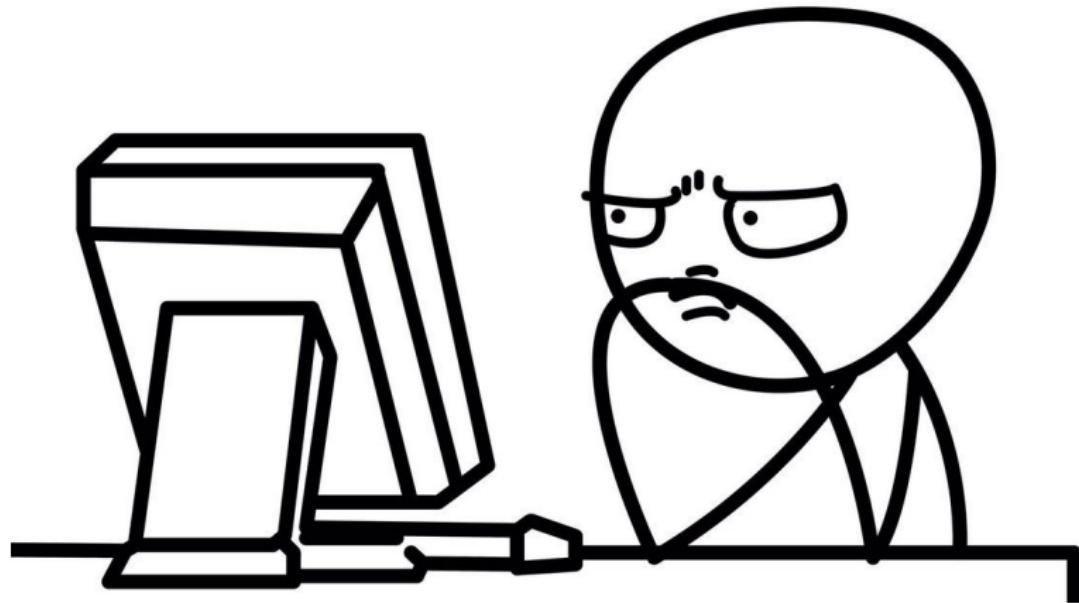
- Иначе, существует константа K такая, что

Zipf's Law

$$\forall f, r : f * r \approx K$$



Демо: закон Ципфа



- Графематический или фонетический анализ

Russian cursive

ИИИИИ ИИИИ
иИиИиИи иИиИи
L i sh i L i L i L i

- Графематический или фонетический анализ
- Токенизация

Russian cursive

Ишиши ишиши

ишиши ишиши

L i sh i L i L i L i i

- Графематический или фонетический анализ
- Токенизация
- Морфологический анализ,
лемматизация

Russian cursive

Ишиши ишиши

и|и|ши|и|и и|и|и|ши|и

L i sh i L i L i L i i

- Графематический или фонетический анализ
 - Токенизация
 - Морфологический анализ,
лемматизация
 - Синтаксический анализ

Russian cursive

Музыка Музыка

L i sh i L i L i L i i

- Графематический или фонетический анализ
 - Токенизация
 - Морфологический анализ,
лемматизация
 - Синтаксический анализ
 - Семантический анализ

Russian cursive

Музыка Музыка

L i sh i L i L i L i i

Лингвистический анализ

Инструменты

- АОТ
- Snowball
- Stemka
- pymorphy
- Myaso
- Eureka Engine
- ISPRAS API Texterra
- pymystem3
- phpmorph
- Pullenti SDK
- FreeLing
- NLTK
- TextBlob
- MBSP
- Pattern
- natural
- MAnalyzer
- hunpos
- SVMTool
- Twitter NLP and Part-of-Speech Tagger
- Stanford Log-linear Part-Of-Speech Tagger
- RussianMorphology
- RussianPOSTagger
- mystem
- TreeTagger
- TnT
- Морфпер
- RCO
- AskNet
- Solarix
- ОРФО
- STARLING
- mystem-scala
- zamgi

- Самый простой способ нормализации слов



- Самый простой способ нормализации слов
- База правил, по которым «отрубаются» суффиксы и окончания



- Самый простой способ нормализации слов
- База правил, по которым «отрубаются» суффиксы и окончания
- Мартин Портер (1980)



- Самый простой способ нормализации слов
- База правил, по которым «отрубаются» суффиксы и окончания
- Мартин Портер (1980)
- Snowball Stemmer



- Самый простой способ нормализации слов
- База правил, по которым «отрубаются» суффиксы и окончания
- Мартин Портер (1980)
- Snowball Stemmer
- **Достоинства:** простота, скорость



- Самый простой способ нормализации слов
- База правил, по которым «отрубаются» суффиксы и окончания
- Мартин Портер (1980)
- Snowball Stemmer
- **Достоинства:** простота, скорость
- **Недостатки:** смотрим на примерах



- Усложнённые алгоритмы



Морфологический анализ

- Усложнённые алгоритмы
- В основе -
морфологический словарь
(чаще всего А.А.Зализняка
или OpenCorpora)



Морфологический анализ

- Усложнённые алгоритмы
- В основе -
морфологический словарь
(чаще всего А.А.Зализняка
или OpenCorpora)
- Предиктивные алгоритмы



Морфологический анализ

- Усложнённые алгоритмы
- В основе -
морфологический словарь
(чаще всего А.А.Зализняка
или OpenCorpora)
- Предиктивные алгоритмы
- **Достоинства:** сложность
реализации, качество
нормализации



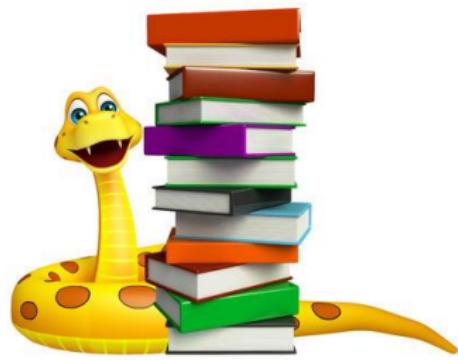
Морфологический анализ

- Усложнённые алгоритмы
- В основе -
морфологический словарь
(чаще всего А.А.Зализняка
или OpenCorpora)
- Предиктивные алгоритмы
- **Достоинства:** сложность
реализации, качество
нормализации
- **Недостатки:** работает
дольше, и всё равно не
идеален



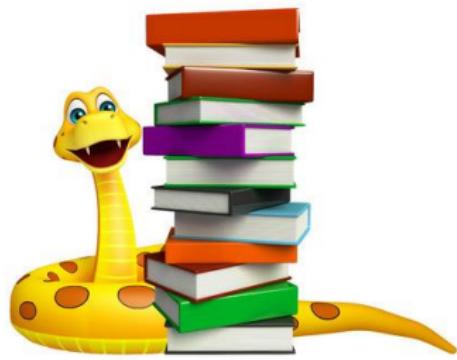
Морфологический анализ PyMorph

- Умеет приводить слово к нормальной форме и ставить слово в нужную форму



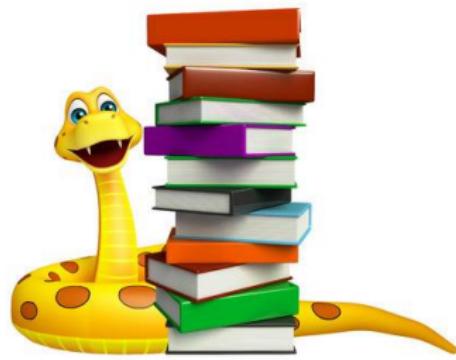
Морфологический анализ PyMorph

- Умеет приводить слово к нормальной форме и ставить слово в нужную форму
- Словарь OpenCorpora



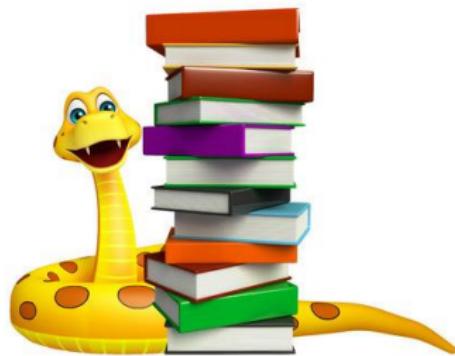
Морфологический анализ PyMorph

- Умеет приводить слово к нормальной форме и ставить слово в нужную форму
- Словарь OpenCorpora
- Продвинутые алгоритмы сжатия информации



Морфологический анализ PyMorph

- Умеет приводить слово к нормальной форме и ставить слово в нужную форму
- Словарь OpenCorpora
- Продвинутые алгоритмы сжатия информации
- Поиск парадигм в словаре



Морфологический анализ MyStem

- Разработка И. Сегаловича



Морфологический анализ MyStem

- Разработка И. Сегаловича
- Умеет снимать морфологическую неоднозначность



Морфологический анализ MyStem

- Разработка И. Сегаловича
- Умеет снимать морфологическую неоднозначность
- Отдельное приложение

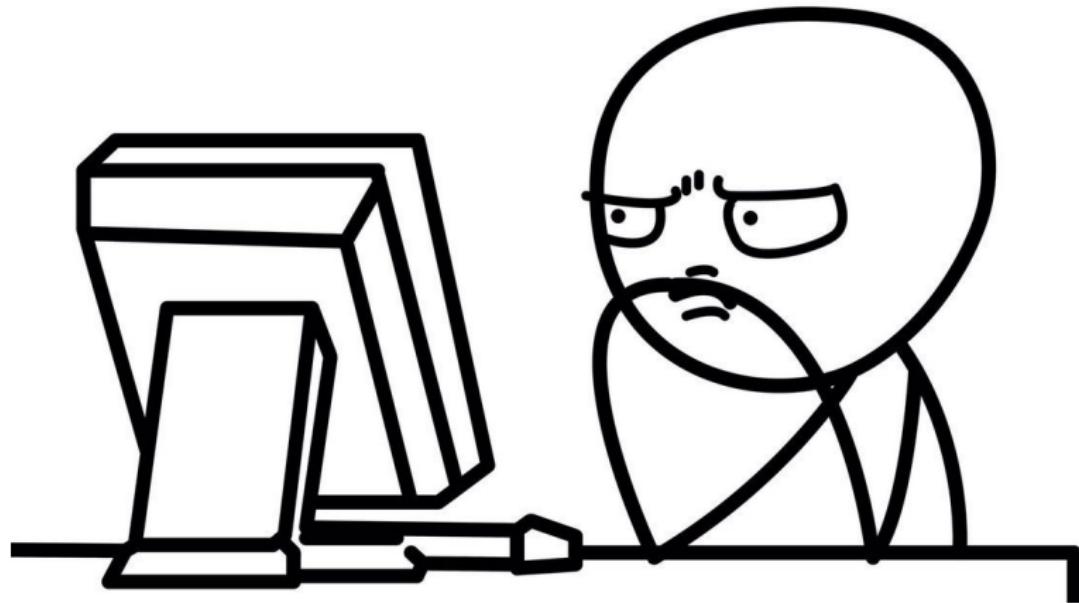


Морфологический анализ MyStem

- Разработка И. Сегаловича
- Умеет снимать морфологическую неоднозначность
- Отдельное приложение
- Интерфейс для Python - pymystem3



Демо: лемматизация



Векторное представление документа

Bag of Words

- Самый простой способ представления документа

Векторное представление документа

Bag of Words

- Самый простой способ представления документа
- Имеем словарь из N слов

Векторное представление документа

Bag of Words

- Самый простой способ представления документа
- Имеем словарь из N слов
- Каждое слово имеет уникальный индекс

Векторное представление документа Bag of Words

- Самый простой способ представления документа
- Имеем словарь из N слов
- Каждое слово имеет уникальный индекс
- Сопоставим документу булев вектор X размерности N , где $X_m \in \{0, 1\}$ - встретилось или нет слово под номером m в документе

Векторное представление документа

Bag of Words

- Самый простой способ представления документа
- Имеем словарь из N слов
- Каждое слово имеет уникальный индекс
- Сопоставим документу булев вектор X размерности N , где $X_m \in \{0, 1\}$ - встретилось или нет слово под номером m в документе

the dog is on the table



Bag Of Words

Векторное представление документа TF-IDF

- Классическая метрика с широким спектром применения

Векторное представление документа TF-IDF

- Классическая метрика с широким спектром применения
- Измеряет значимость слова в контексте документа, являющегося частью некоторого корпуса

Векторное представление документа TF-IDF

- Классическая метрика с широким спектром применения
- Измеряет значимость слова в контексте документа, являющегося частью некоторого корпуса
- **TF:** (term frequency) - мера частотности слова в документе

Векторное представление документа TF-IDF

- Классическая метрика с широким спектром применения
- Измеряет значимость слова в контексте документа, являющегося частью некоторого корпуса
- **TF:** (term frequency) - мера частотности слова в документе
- Характеризует значимость слова в рамках данного документа

Векторное представление документа TF-IDF

- Классическая метрика с широким спектром применения
- Измеряет значимость слова в контексте документа, являющегося частью некоторого корпуса
- **TF:** (term frequency) - мера частотности слова в документе
- Характеризует значимость слова в рамках данного документа
- **IDF:** (inverse document frequency) - мера "редкости" слова в масштабах всего корпуса

Векторное представление документа

TF-IDF

- Классическая метрика с широким спектром применения
- Измеряет значимость слова в контексте документа, являющегося частью некоторого корпуса
- **TF:** (term frequency) - мера частотности слова в документе
- Характеризует значимость слова в рамках данного документа
- **IDF:** (inverse document frequency) - мера "редкости" слова в масштабах всего корпуса
- Результат эвристического наблюдения - редкие слова несут больше смысла

Векторное представление документа TF-IDF

- Классическая метрика с широким спектром применения
- Измеряет значимость слова в контексте документа, являющегося частью некоторого корпуса
- **TF:** (term frequency) - мера частотности слова в документе
- Характеризует значимость слова в рамках данного документа
- **IDF:** (inverse document frequency) - мера "редкости" слова в масштабах всего корпуса
- Результат эвристического наблюдения - редкие слова несут больше смысла
- Могут вычисляться по-разному

Векторное представление документа

TF-IDF: как вычислять TF?

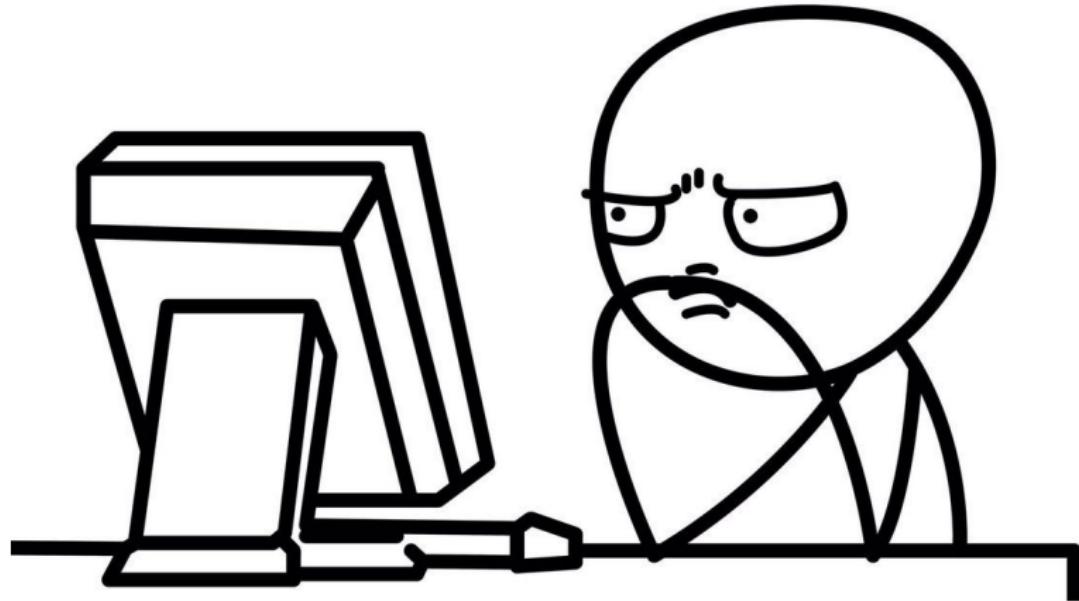
Название	Формула
Частота	$f_{t,d}$
Нормированная частота	$\frac{f_{t,d}}{\sum_k f_{t_k,d}}$
Бинарная встречаемость	$\begin{cases} 0, & f_{t,d} = 0 \\ 1, & f_{t,d} \geq 1 \end{cases}$
Логарифм частоты	$\log(1 + f_{t,d})$
Double normalization	$K + (1 - K) \frac{f_{t,d}}{\max(f_{t_i,d})}$

Векторное представление документа

TF-IDF: как вычислять IDF?

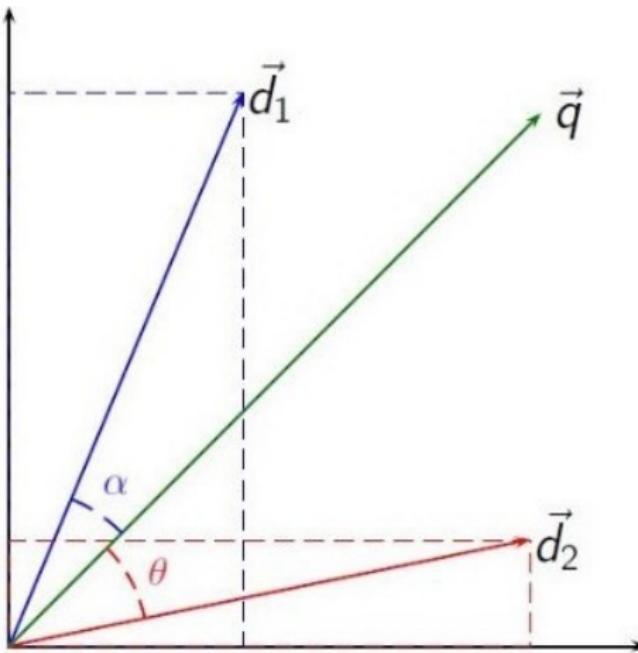
Название	Формула
Стандарт	$\log \frac{N}{n_t}$
«Сглаженный»	$\log(1 + \frac{N}{n_t})$
Сглаженный + $\max(DF)$	$\log \left(1 + \frac{\max_t n_t}{n_t}\right)$
Вероятностная обратная встречааемость	$\log \frac{N - n_t}{n_t}$

Демо: TF-IDF

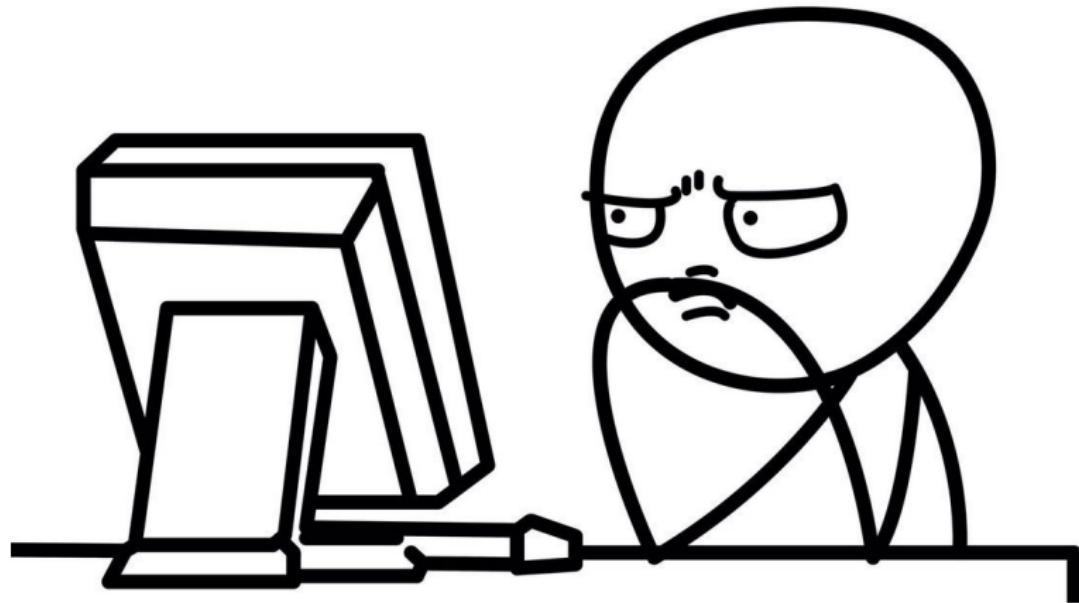


Косинусная близость

$$1 - \text{similarity}(d_2, q) = \cos(\theta) = \frac{d_2 \cdot q}{\|d_2\| \|q\|} = \frac{\sum_{i=1}^n d_{2i} \times q_i}{\sqrt{\sum_{i=1}^n (d_{2i})^2} \times \sqrt{\sum_{i=1}^n (q_i)^2}}$$



Демо: Косинусная близость, поиск похожих документов



Векторное представление документа

Bag of Words: недостатки

- Высокая размерность признакового пространства

Векторное представление документа

Bag of Words: недостатки

- Высокая размерность признакового пространства
- Игнорируется информация о семантической связи слов

Векторное представление документа

Bag of Words: недостатки

- Высокая размерность признакового пространства
- Игнорируется информация о семантической связи слов
- Теряется информация о порядке слов

Object → Bag of ‘words’



Векторное представление документа

Bag of Words: недостатки

- Высокая размерность признакового пространства
- Игнорируется информация о семантической связи слов
- Теряется информация о порядке слов
- Нельзя учесть информацию из неразмеченных документов

Object → Bag of ‘words’



Векторное представление документа Word2Vec

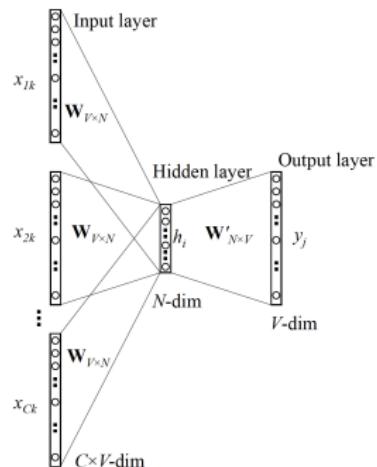
- **Дистрибутивная гипотеза:**
лингвистические единицы,
встречающиеся в схожих
контекстах, имеют близкие
значения.

Векторное представление документа Word2Vec

- **Дистрибутивная гипотеза:**
лингвистические единицы,
встречающиеся в схожих
контекстах, имеют близкие
значения.
- **Автоэнкодер:** специальная
архитектура нейронной сети,
направленная на получение на
выходном слое отклика, наиболее
близкого к входному вектору

Векторное представление документа Word2Vec

- **Дистрибутивная гипотеза:**
лингвистические единицы,
встречающиеся в схожих
контекстах, имеют близкие
значения.
- **Автоэнкодер:** специальная
архитектура нейронной сети,
направленная на получение на
выходном слое отклика, наиболее
близкого к входному вектору
- Будем предсказывать слово по
его контексту (CBOW) или
контекст слова по нему самому
(Skip-Gram)



Векторное представление документа

Word2Vec: что получаем

- Можем учитывать неограниченный объём неразмеченных текстов

Векторное представление документа

Word2Vec: что получаем

- Можем учитывать неограниченный объём неразмеченных текстов
- Векторы похожих по смыслу слов близки в семантическом пространстве - можем находить похожие слова

Векторное представление документа

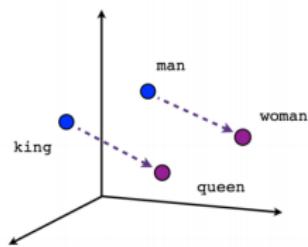
Word2Vec: что получаем

- Можем учитывать неограниченный объём неразмеченных текстов
- Векторы похожих по смыслу слов близки в семантическом пространстве - можем находить похожие слова
- Можно решать семантические пропорции

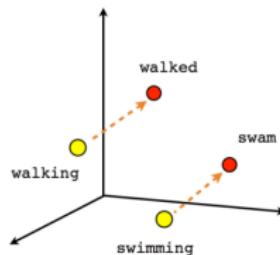
Векторное представление документа

Word2Vec: что получаем

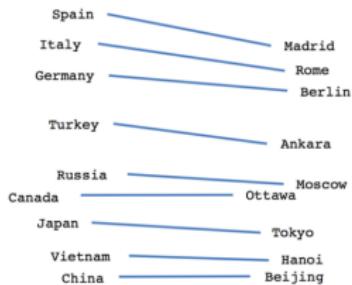
- Можем учитывать неограниченный объём неразмеченных текстов
- Векторы похожих по смыслу слов близки в семантическом пространстве - можем находить похожие слова
- Можно решать семантические пропорции



Male-Female



Verb tense



Country-Capital

Векторное представление документа

Word2Vec: недостатки

- Не умеет различать синонимы и антонимы

Векторное представление документа

Word2Vec: недостатки

- Не умеет различать синонимы и антонимы
- Непонятно, как перейти от векторного представления слов к документам

Векторное представление документа

Word2Vec: недостатки

- Не умеет различать синонимы и антонимы
- Непонятно, как перейти от векторного представления слов к документам
- Не решает проблем омонимии и анафоры

Векторное представление документа

Word2Vec: недостатки

- Не умеет различать синонимы и антонимы
- Непонятно, как перейти от векторного представления слов к документам
- Не решает проблем омонимии и анафоры
- Близость слов в тексте не обязательно тождественна семантической близости

Векторное представление документа

Word2Vec: недостатки

- Не умеет различать синонимы и антонимы
- Непонятно, как перейти от векторного представления слов к документам
- Не решает проблем омонимии и анафоры
- Близость слов в тексте не обязательно тождественна семантической близости
- Не умеет работать с незнакомыми словами

Список литературы

Что читать и куда гуглить

- Маннинг К.Д., Рагхаван П., Шютце Х. "Введение в информационный поиск" - хорошо дана база, *Bag of Words, TF-IDF*

Список литературы

Что читать и куда гуглить

- Маннинг К.Д., Рагхаван П., Шютце Х. "Введение в информационный поиск" - хорошо дана база, *Bag of Words, TF-IDF*
- С. Николенко, А. Кадурин, Е. Архангельская "Глубокое обучение. Погружение в мир нейронных сетей" - книга в целом про более продвинутые вещи, но есть отличная глава про *Word2Vec*

Список литературы

Что читать и куда гуглить

- Маннинг К.Д., Рагхаван П., Шютце Х. "Введение в информационный поиск" - хорошо дана база, *Bag of Words, TF-IDF*
- С. Николенко, А. Кадурин, Е. Архангельская "Глубокое обучение. Погружение в мир нейронных сетей" - книга в целом про более продвинутые вещи, но есть отличная глава про *Word2Vec*
- Web-scraping - обратите внимание на пакеты scrapy и selenium

Список литературы

Что читать и куда гуглить

- Маннинг К.Д., Рагхаван П., Шютце Х. "Введение в информационный поиск" - хорошо дана база, *Bag of Words, TF-IDF*
- С. Николенко, А. Кадурин, Е. Архангельская "Глубокое обучение. Погружение в мир нейронных сетей" - книга в целом про более продвинутые вещи, но есть отличная глава про *Word2Vec*
- Web-scraping - обратите внимание на пакеты scrapy и selenium
- <http://nlpublish.ru> - русскоязычный wiki по NLP

Список литературы

Что читать и куда гуглить

- Маннинг К.Д., Рагхаван П., Шютце Х. "Введение в информационный поиск" - хорошо дана база, *Bag of Words, TF-IDF*
- С. Николенко, А. Кадурин, Е. Архангельская "Глубокое обучение. Погружение в мир нейронных сетей" - книга в целом про более продвинутые вещи, но есть отличная глава про *Word2Vec*
- Web-scraping - обратите внимание на пакеты scrapy и selenium
- <http://nlpublish.ru> - русскоязычный wiki по NLP
- <http://www.dialog-21.ru> - сайт конференции Диалог, с материалами и датасетами к соревнованиям