Компьютерная лингвистика (знакомство)

НИУ ВШЭ, магистратура “Журналистика данных”
15 января 2019г.

О курсе

Практический курс. Будут задания с элементами программирования на Python.

Отчетность
Домашние задания (ДЗ), в спорных случаях возможен экзамен

Дедлайны
ДЗ вовремя — полный балл
ДЗ на неделю позже — минус 1 балл
все, что позже, не оценивается

Что такое компьютерная лингвистика (КЛ)?
Чем она занимается?

Немного терминологии

Computational linguistics

Natural language processing (NLP)

Language technology

Artificial Intelligence

Данные

Языковые данные: текст, звучащая речь

Неструктурированные данные

Текст/речь — закодированная на естественном языке информация. Человек знает язык и может понимать тексты/речь.
Компьютер “понимает” двоичный код. Компьютер “склонен” считать, а не читать.

КЛ вокруг нас

спеллчекер, автозамена

КЛ вокруг нас

спеллчекер, автозамена

спам-фильтры

КЛ вокруг нас

спеллчекер, автозамена

спам-фильтры

машинный перевод

КЛ вокруг нас

спеллчекер, автозамена

спам-фильтры

машинный перевод

распознавание речи

КЛ вокруг нас

спеллчекер, автозамена

спам-фильтры

машинный перевод

распознавание речи

чат-боты, голосовые помощники

КЛ вокруг нас

спеллчекер, автозамена

спам-фильтры

машинный перевод

распознавание речи

чат-боты, голосовые помощники

информационный поиск

КЛ, о которой знают специалисты

Бывают “удобные” pdf-файлы (например, со сканом текста), а бывают “неудобные”. О чем идет речь?

КЛ, о которой знают специалисты

Бывают “удобные” pdf-файлы (например, со сканом текста), а бывают “неудобные”. О чем идет речь?

Ответ: распознавание символов (OCR, optical character recognition)

КЛ, о которой знают специалисты

1) Нужно сосчитать в тексте количество глаголов и прилагательных
2) Нужно найти в тексте все упоминания глагола *быть*.

Какая технология КЛ поможет?

КЛ, о которой знают специалисты

1) Нужно сосчитать в тексте количество глаголов и прилагательных
2) Нужно найти в тексте все упоминания глагола *быть*.

Какая технология КЛ поможет?

Ответ: морфологический анализатор

КЛ, о которой знают специалисты

Хотим сократить текст, убрав из него все причастные и деепричастные обороты.

КЛ, о которой знают специалисты

Хотим сократить текст, убрав из него все причастные и деепричастные обороты.

Ответ: синтаксический анализ

КЛ, о которой знают специалисты

Хотим проанализировать комментарии к одной из нашумевших новостей и оценить отношение к событию.

КЛ, о которой знают специалисты

Хотим проанализировать комментарии к одной из нашумевших новостей и оценить отношение к событию.

Ответ: анализ тональности/сентимент анализ (sentiment analysis)

КЛ, о которой знают специалисты

Хотим быстро проанализировать массив текстов и понять, о чем они.

КЛ, о которой знают специалисты

Хотим быстро проанализировать массив текстов и понять, о чем они.

Ответ: извлечение ключевых слов

КЛ, о которой знают специалисты

Реферирование текстов
Извлечение именованных сущностей
Анализ жанров
Поиск плагиата
Генерация текстов

Характер задач

Анализ (большинство)

Синтез

И то, и другое

NLP meets journalism

Detection of biased language

Определение, “перекоса” в подаче информации, для новостей, которые могут быть написаны с позиции одной из конфликтующих сторон, например, пророссийская новость или проукраинская.

На основе лексического состава, а также знании о источнике новости,

NLP meets journalism

Fake news detection

Является новость фейковой или нет.

Определяется на основе лексического состава, риторических структур.

Проблема — сложно сформулировать признаки, плюс не всегда человек сам может отличить фейк от нефейка.

NLP meets journalism

Trend prediction

Предсказание того, какого характера новости выйдут, на основе знания о том, какими были тематики новостей ранее.

NLP meets journalism

Trend prediction

Предсказание того, какого характера новости выйдут, на основе знания о том, какими были тематики новостей ранее.

Кластеризация по тематикам

Выделения тематик в массиве текстовых данных.
На основе анализа лексического содержания текстов.

Сложности при работе с текстовыми данными

Например, при поиске

1. Ищем в Google/Yandex поисковике сочетание “*Екатерина Великая*”, какие упоминания в ссылках находим? А если делаем поиск по странице?

2. Какие проблемы будут с запросами apple или вышка?

3. Какие еще упоминания найдутся по запросу “телефон”?

4. Что не так с выражением “студенты из Львова поехали в Киев”?

a с “flying planes can be dangerous”?

5. Простой - простой, дорога - дорога, sport - sport

Сложности при работе с текстовыми данными

Например, при поиске

1. Ищем в Google/Yandex поисковике сочетание “*Екатерина Великая*”, какие упоминания в ссылках находим? А если делаем поиск по странице?

2. Какие проблемы будут с запросами apple или вышка?

3. Какие еще упоминания найдутся по запросу “телефон”?

4. Что не так с выражением “студенты из Львова поехали в Киев”?

a с “flying planes can be dangerous”?

5. Простой - простой, дорога - дорога, sport - sport

Сложности при работе с текстовыми данными

Например, при поиске

1. Ищем в Google/Yandex поисковике сочетание “*Екатерина Великая*”, какие упоминания в ссылках находим? А если делаем поиск по странице?

2. Какие проблемы будут с запросами apple или вышка?

3. Какие еще упоминания найдутся по запросу “телефон”?

4. Что не так с выражением “студенты из Львова поехали в Киев”?

a с “flying planes can be dangerous”?

5. Простой - простой, дорога - дорога, sport - sport

Сложности при работе с текстовыми данными

Например, при поиске

1. Ищем в Google/Yandex поисковике сочетание “*Екатерина Великая*”, какие упоминания в ссылках находим? А если делаем поиск по странице?

2. Какие проблемы будут с запросами apple или вышка?

3. Какие еще упоминания найдутся по запросу “телефон”?

4. Что не так с выражением “студенты из Львова поехали в Киев”?

a с “flying planes can be dangerous”?

5. Простой - простой, дорога - дорога, sport - sport

Сложности при работе с текстовыми данными

Например, при поиске

1. Ищем в Google/Yandex поисковике сочетание “*Екатерина Великая*”, какие упоминания в ссылках находим? А если делаем поиск по странице?

2. Какие проблемы будут с запросами apple или вышка?

3. Какие еще упоминания найдутся по запросу “телефон”?

4. Что не так с выражением “студенты из Львова поехали в Киев”?

a с “flying planes can be dangerous”?

5. Простой - простой, дорога - дорога, sport - sport

Сложности при работе с текстовыми данными

Например, при поиске

1. Ищем в Google/Yandex поисковике сочетание “*Екатерина Великая*”, какие упоминания в ссылках находим? А если делаем поиск по странице?

2. Какие проблемы будут с запросами apple или вышка?

3. Какие еще упоминания найдутся по запросу “телефон”?

4. Что не так с выражением “студенты из Львова поехали в Киев”?

a с “flying planes can be dangerous”?

5. Простой - простой, дорога - дорога, sport - sport

Неоднозначность (омонимия)

Лексическая

Морфологическая

Лексико-морфологическая

Синтаксическая

Sum up

КЛ имеет дело с языковыми данными (текстом/речью)

Это неструктурированные данные, отсюда сложности в работе с ними

КЛ для КЛ: морфологические анализаторы, синтаксические анализаторы

Существует множество КЛ технологий, которые нам очень хорошо знакомы

Есть более специальные области КЛ (какие?)

Есть задачи, которые могут быть интересны журналистам