# Информационные технологии в прикладных системах

## Лекция 1: Извлечение информации из Интернет-ресурсов

Ключевые моменты: работа с более «интеллектуальными» заданиями, задания тематического моделирования

Интеллектуальная обработка: определить проблему -> определить полученные данные -> приведение данных к необходимому виду (предобработка) -> моделирование данных -> тестирование на выборке -> проверка и развертывание, визуализация

Результат обработки – информационный продукт

(идея: анализировать тренды с показов одежды)

Web-Mining

использование методов интеллектуального анализа данных для:

- выявление общих закономерностей

- извлечение информации из веб-ресурсов

- автоматического обнаружения веб-документов

- решения задач бизнес-аналитики

Парсинг

PHP: Simple HTML DOM Parser, PHPQuery

Python: **BeautifulSoup,** lxml, html5lib

Regular expressions

## Лекция 2: Предобработка текста (preprocessing)

**Natural Language Processing** (NLP) – использование математических моделей для описания лингвистических закономерностей.

1. Применение известных математических подходов для выявления закономерностей языка.
2. Создание моделей для решения практических задач и разработка программ.

Подходы к анализу текста:

* Формально-логический (при помощи правил и формальной логики)
* Стохастический (вероятностный)

Основные понятия NLP:

* Токен – цепочка любых символов, находящаяся в тексте между двумя знаками пробела.
* Лексема – слово, имеющая определенное лексическое значение.
* Словоформа – лексема с определенными грамматическими параметрами: падеж, число, время и т.п.
* Предложение - цепочка токенов между двумя знаками конца предложения
* Текст – линейная последовательность токенов.
* Корпус – набор текстов, используется для обучения моделей, проверки гипотез и статистического лингвистического анализа
* Bag of words (BOW) – это репрезентативная модель, используемая для упрощения обработки текста. Не учитывает порядок слов. Главная задача – определение количества вхождений слов в данный текст.

N-граммы

– наборы из N подряд идущих токенов.

Примеры:

* Униграммы: наборы, подряд, идущих, токенов
* Биграммы: наборы подряд, подряд идущих, идущих токенов
* Триграммы: наборы подряд идущих, подряд идущих токенов

### Количественные характеристики

Абсолютная частота вхождения токена – число найденных в тексте вхождений токена.

Абсолютная частота лексемы (слова) – сумма зафиксированных в тексте всех словоформ лексемы.

Относительная частота вхождения лексемы (слова) – сумма зафиксированных в тексте вхождений всех словоформ лексемы, поделенная на общее число токенов в тексте (длина текста).

### Предварительная обработка текста

* Замена множественных пробелов на одинарный
* Разбиение на токены
* Удаление знаков препинания и спецсимволов (опционально)
* Перевод на нижний регистр
* Нормализация

– стемминг,

– лемматизация

* Разметка
* Подсчет вхождений

Стемминг – это процесс нахождения основы слова для заданного исходного слова.

Основа слова – неизменяемая часть слова, которая выражает его лексическое значение.

Основа слова необязательно совпадает с морфологическим корнем слова. Применяется в поисковых системах для обобщения поискового запроса пользователя.

**Алгоритмы стемматизации** или **стеммер**

Лемматизация – замена слова его смысловой канонический формой (инфинитив для глагола, именительный падеж единственного числа – для сущ. и прил.)

Примеры:

ценами – цена

### Библиотеки

* MyStem
* nltk
* **genism**
* spacy
* ***pymorphy2***

### Дополнительные опции предобработки

Вид токена:

* токен: одно слово
* n-грамма – последовательность из n токенов (униграмма, биграмма и т.д.)
* предложение

Разметка термов

- некоторая характеристика или набор характеристик терма.

Пример: если токен - слово, то разметка: часть речи

Результат предобработки

* Словарь нормализованных слов или n-грамм (термы)
* Частота вхождения каждого терма
* Разметка терма

POS – Part of Speech

## Лекция 3: Vector Space Model (VSM)

* алгебраическая модель для представления текстовых документов в виде векторов из численных весов или коэффициентов
* представление каждого документа коллекции в качестве точки в многомерном пространстве
* автоматически извлекает полезные знания из заданного текста, однако определяет его смысл с сильным упрощением
* является основой для решения многих задач информационного поиска: поиск документа по запросу, классификация документов, кластеризация документов. Сопоставление текстов.

(Алгоритм классификации, исключать стоп-слова: «и,в,и тд.» )

### Математическая обработка

* Получение частот из корпусов
* Назначение весов термов
* Сглаживание
* Вычисление меры сходства векторов

### Веса термов

* булевский вес – равен 1, если терм встречается в документе и 0 в противном случае
* TF(term frequency- частота терма)
* TF-IDF(term frequency-inverse document frequency – частота терма – обратная частота документа)

TF = ni/Nd

IDF = log(N/di)

* PMI – взаимная встречаемость термов, контекст.

### Относительная частота TF

Чем выше слово – тем реже оно встречается

Косинусная мера сходства текстов

Мера Jaccard

**Мера сходства считается только по 2 объектам.**

### Сжатие векторов

* С большими корпусами мы получим миллионы измерений (осей, контекстов)
* Но вектора очень разреженные, большинство компонентов равны нулю
* Разными математическими трюками можно снизить размерность векторов до разумных значений и всё еще сохранить значимые отношения между ними.
* Такие сжатые вектора называются «word embedding»
* Размерность можно сокращать пока на дойдем до 2 или 3 компонентов. 2-мерное или 3-мерное пространство уже можно визуализировать.   
  PCA, t-SNE

Подходы к построению WE

1. Матрицы совместной встречаемости, факторизованные через SVD
2. Предсказательные модели

Счетные модели

1. посчитать полную матрицу совместной встречаемости

### Предсказательные модели

Применение: word2vec

Алгоритмы: Bag-of-Words, Skipgram

### Построение модели

1. Сначала генерируются случайные вектора для каждого слова
2. Во время обучения мы движемся по корпусу скользящим окном
3. Каждое слово в тексте воспринимается как задача предсказания: мы хотим предсказать это слово при помощи слов-соседей
4. Исход предсказаний (ПРАВИЛНО ИЛИ НЕТ)

<https://ronxin.github.io/wevi/>

Rusvectores.org

nltk (SentimentAnalyzer) – библиотека сентиментных слов