АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ

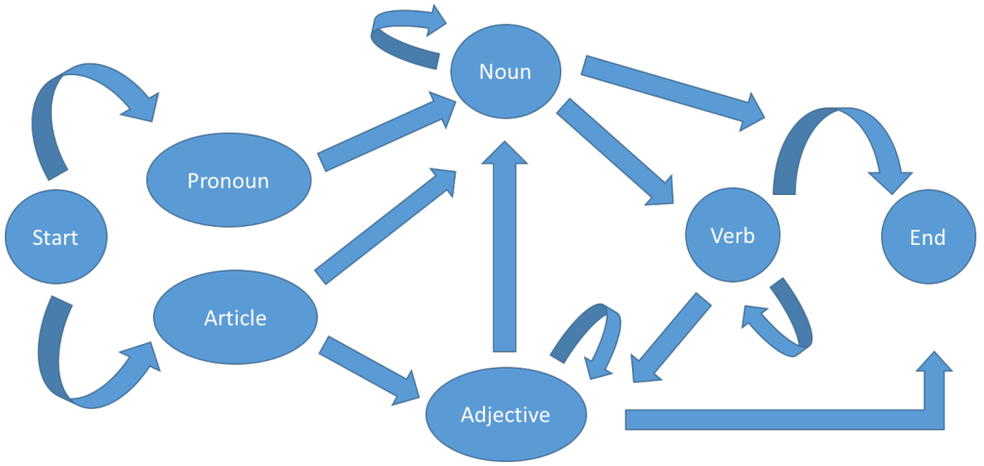
Дмитрий Ильвовский, к.т.н
Екатерина Черняк, к.т.н.

Департамент анализа данных и искусственного интеллекта
Факультет компьютерных наук
НИУ ВШЭ

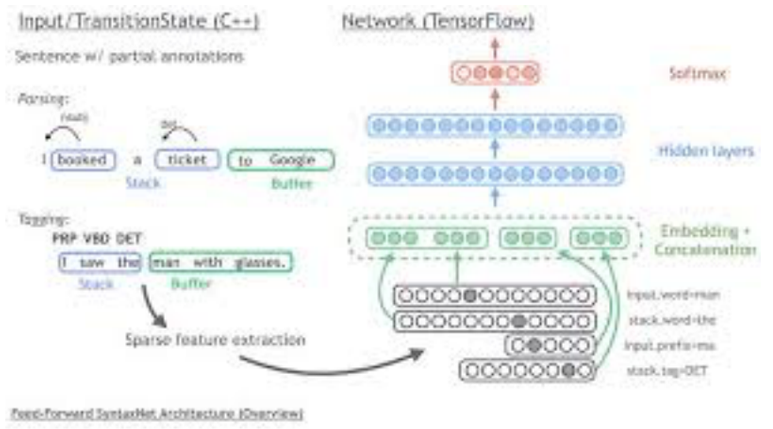
- Темы
- Инструменты
- Источники
- Формы контроля: 4 домашних задания, эссе, зачет
- Курсы в ВШЭ:
  - Машинное обучение для лингвистов
  - Анализ неструктурированных данных

ОБРАБОТКА ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

токенизация
морфология
синтаксис
извлечение ключевых слов и словосочетаний

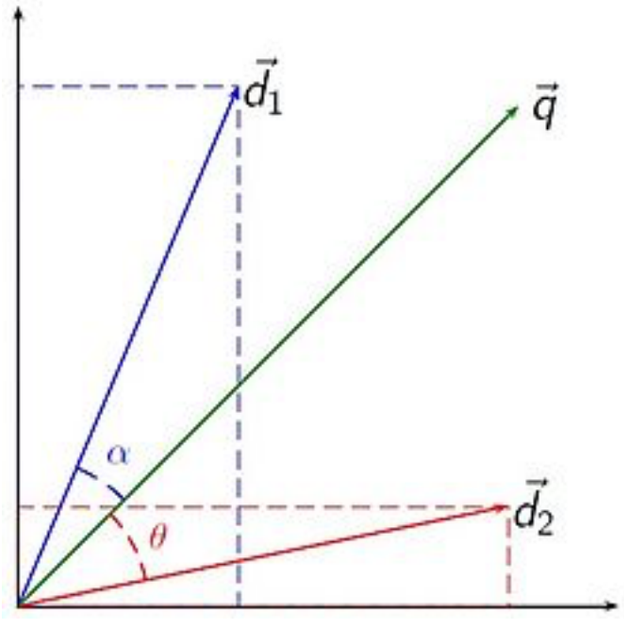


Источник: http://alankarkotwal.github.io/portfolio/pos.html

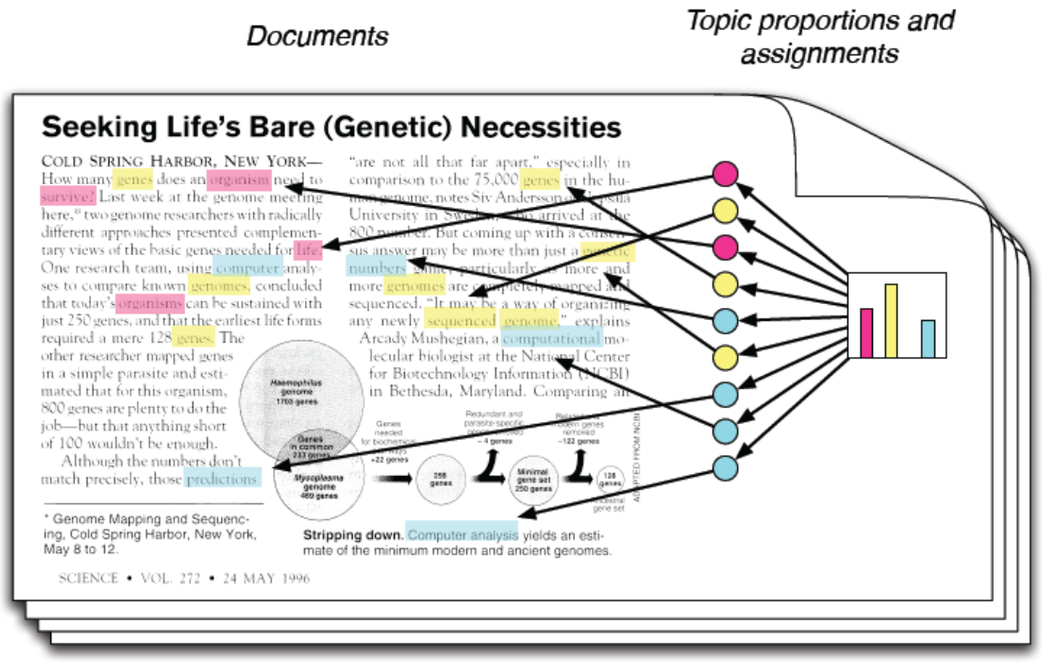


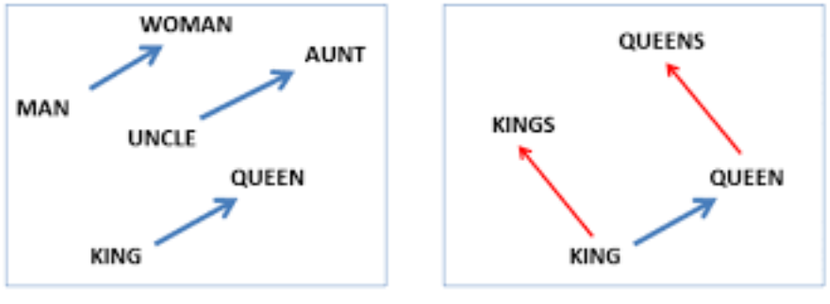
Источник: https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

ВЕКТОРНАЯ МОДЕЛЬ



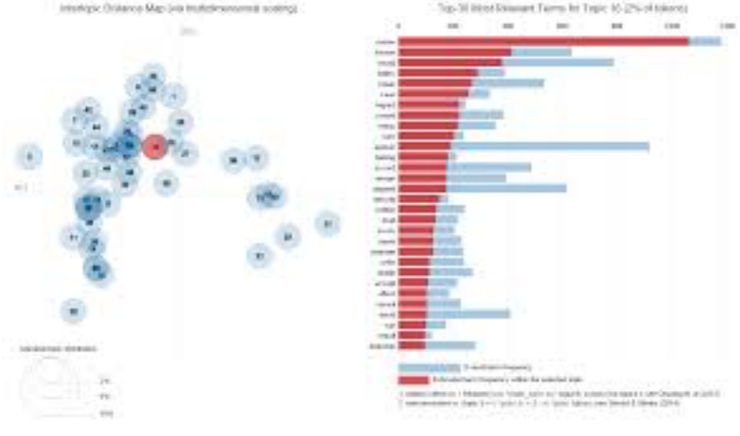
Источник: Wiki





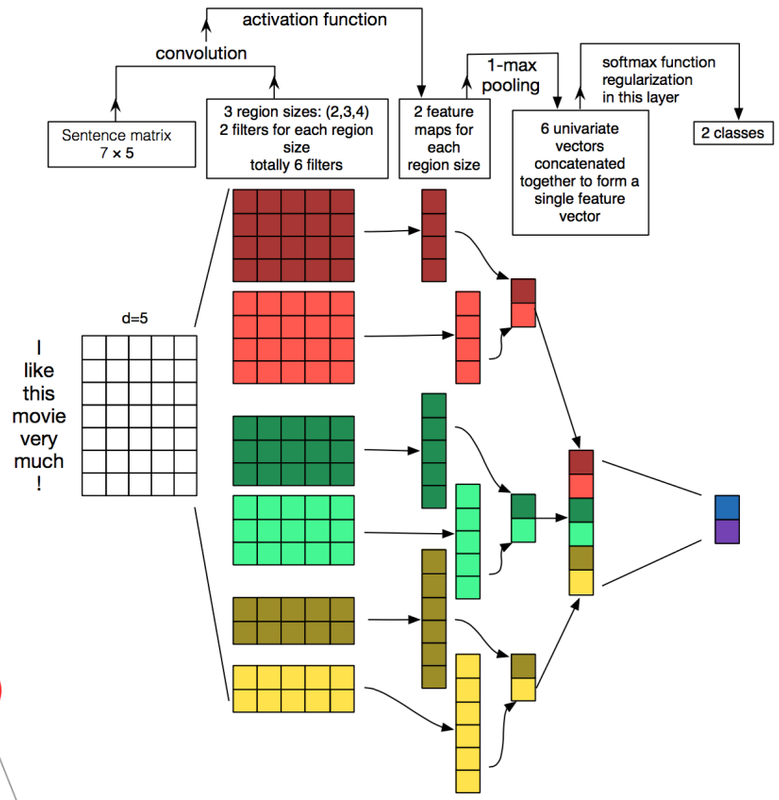
(Mikolov et al., NAACL HLT, 2013)

поиск в векторной модели
снижение размерности в векторной модели
модели скрытых тем
дистрибутивная семантика

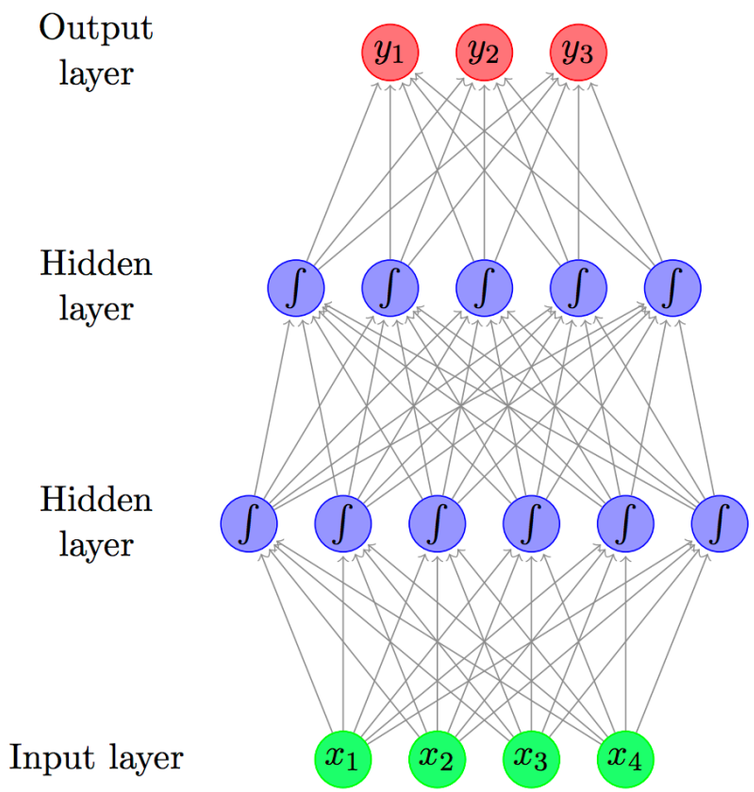


КЛАССИФИКАЦИЯ ТЕКСТОВ

- Метод Наивного Байеса
- Логистическая регрессия
- FastText
- Сети прямого распространения
- Сверточные сети



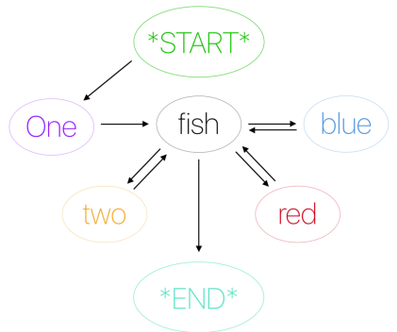
Источник:Kim, 2014



Источник: Goldberg, 2016

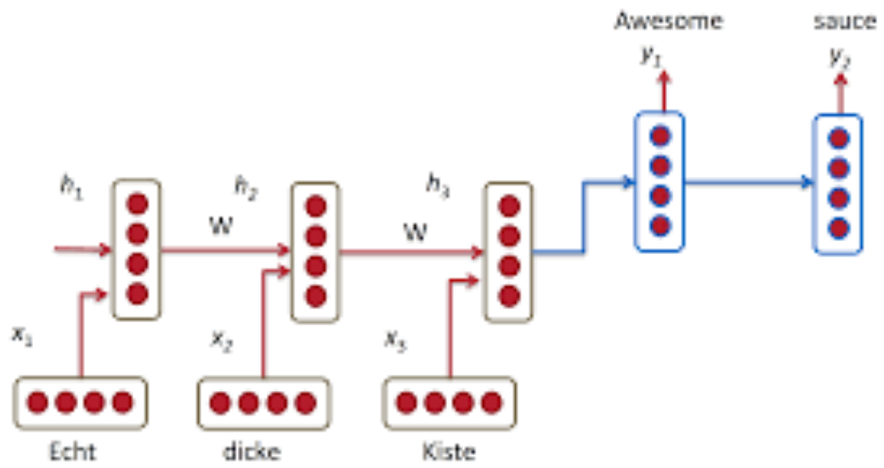
ЯЗЫКОВЫЕ МОДЕЛИ [LANGUAGE MODEL]

- Цепи Маркова
- Нейронные языковые модели
- Рекуррентные нейронные языковые модели
  - LSTM, GRU

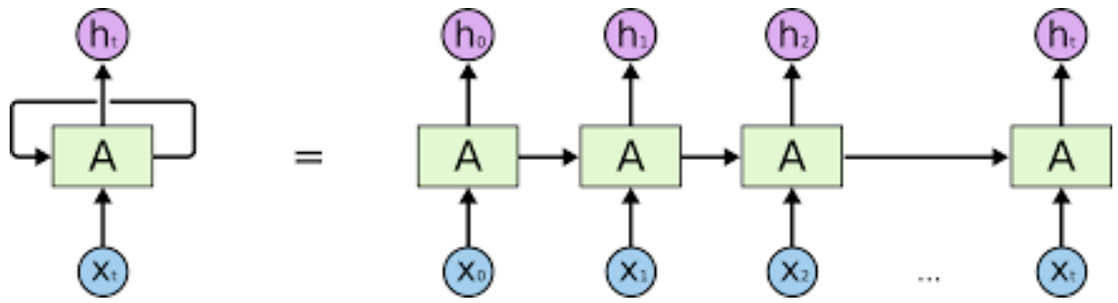


Истоачник: medium.com

- Seq2seq архитектуры



Истоачник: wildml.com



Истоачник: colah.github.com

КЛАССИФИКАЦИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ [SEQUENCE LABELLING]

- Условные случайные поля
- Рекуррентные нейронные сети

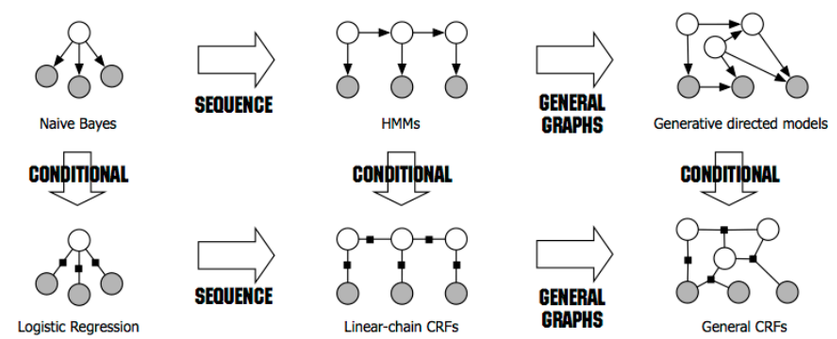


Figure 1.2 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs.

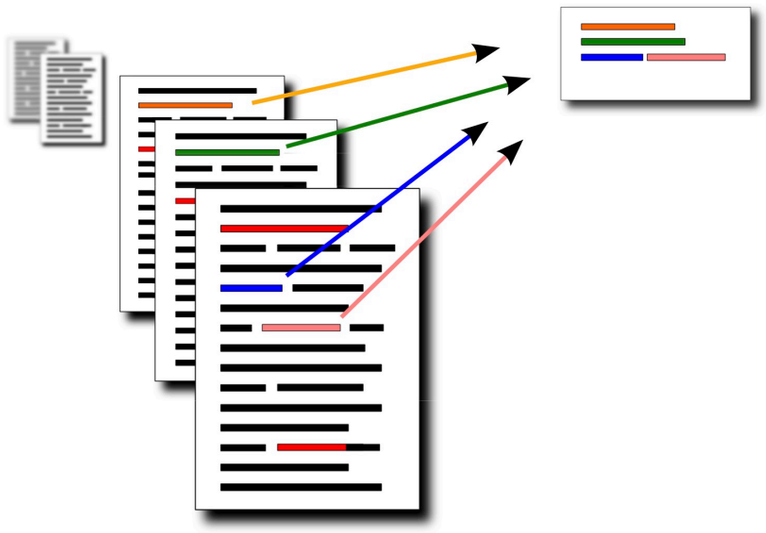
Источник: McCallum, 2012

Извлечение именованных сущностей

И МНОГОЕ ДРУГОЕ

Источник: http://magazine.utoronto.ca

машинный перевод
суммаризация текстов
генерация подписи к изображению
поиск опечаток
распознавание речи
кластеризация текстов
поиск дубликатов
символьные модели
генерация текстов



Источник: http://www.cse.chalmers.se/

Источник: http://copia.com.au/

ИНСТРУМЕНТЫ

- NLTK
- Gensim
- Keras
- PyMorphy2
- PyMystem3
- SyntaxNet
- Томита-парсер

ИСТОЧНИКИ

- Manning, Christopher D., and Hinrich Schütze. Foundations of statistical natural language processing. MIT press, 1999.
- Jurafsky, Daniel. Speech and language processing: An introduction to natural language processing. Computational linguistics, and speech recognition, 2000.
- Goldberg, Yoav. "Neural network methods for natural language processing." *Synthesis Lectures on Human Language Technologies* 10, no. 1 (2017): 1-309.
- Perkins, Jacob. Python text processing with NLTK 2.0 cookbook. Packt Publishing Ltd, 2010.