

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

.....

.....



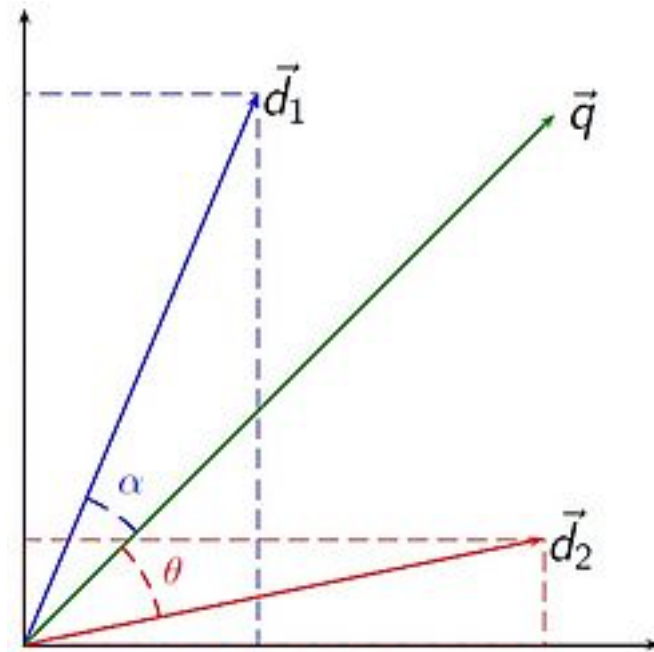
.....



.....



ВЕКТОРНАЯ МОДЕЛЬ



Источник: Wiki

Documents

Topic proportions and
assignments

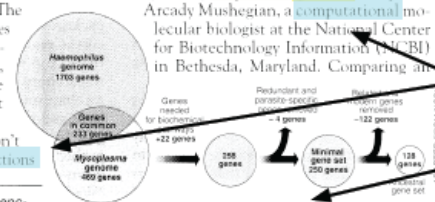
Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson, a biologist at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

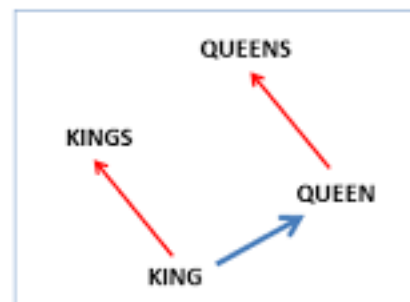
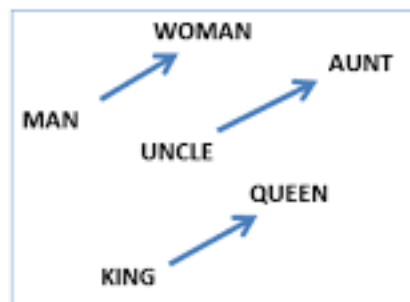
Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

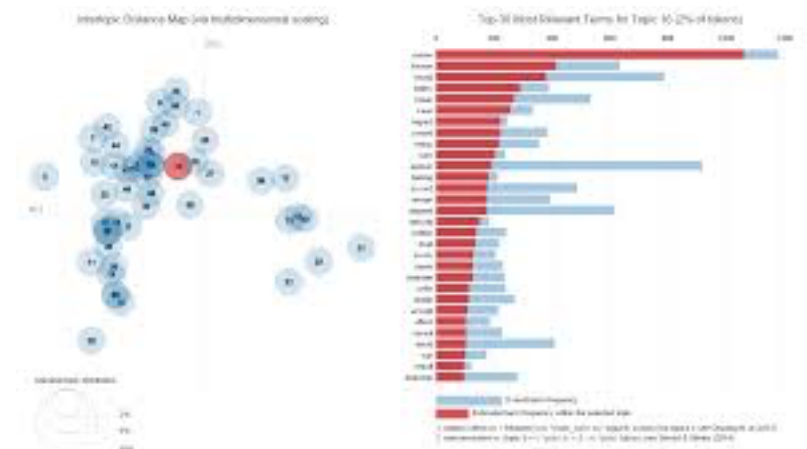
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996



(Mikolov et al., NAACL HLT, 2013)

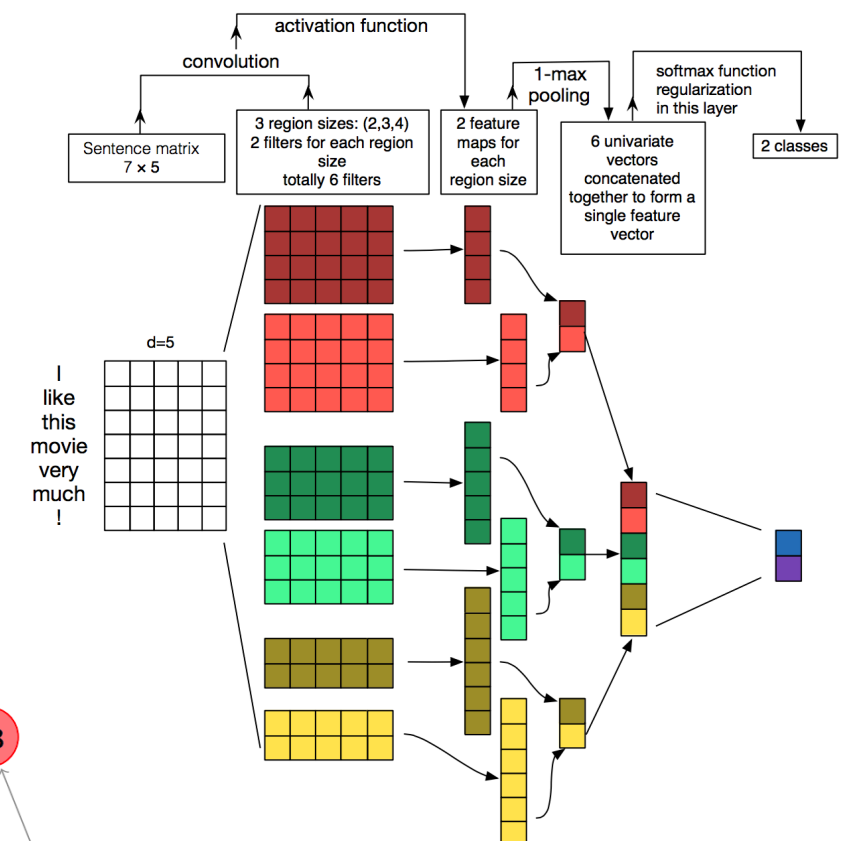
поиск в векторной модели
снижение размерности в векторной модели
модели скрытых тем
дистрибутивная семантика



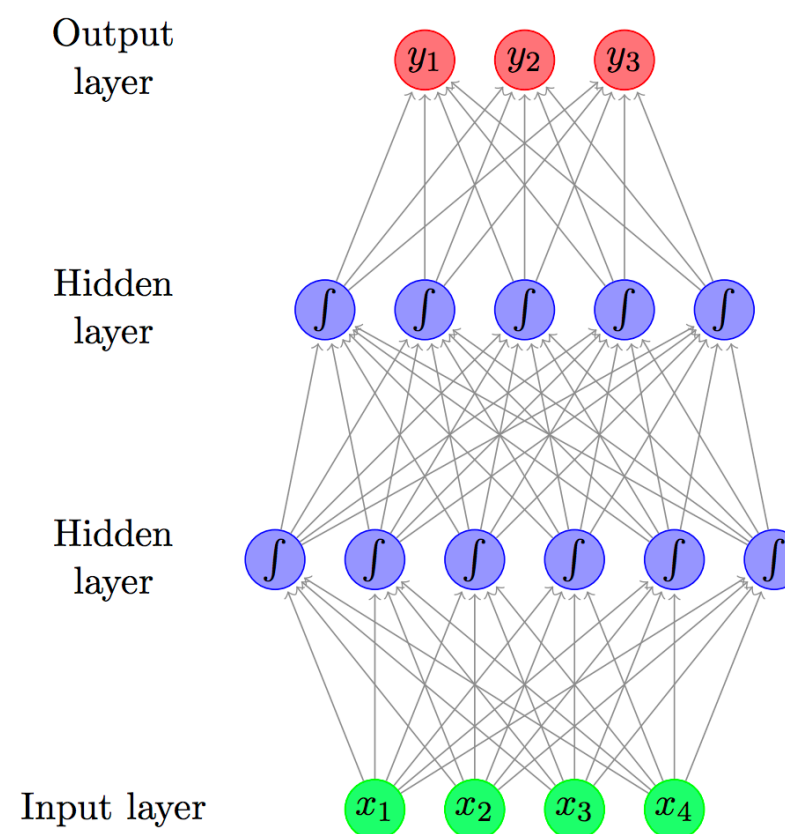
КЛАССИФИКАЦИЯ ТЕКСТОВ

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

*fast*Text



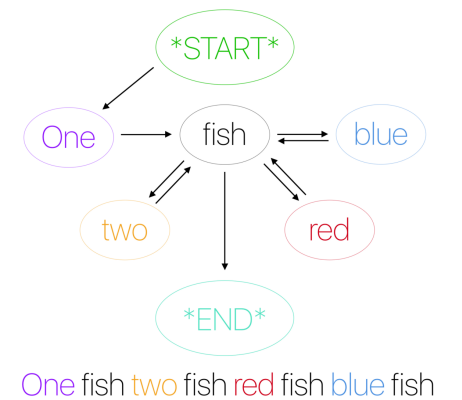
Источник: Kim, 2014



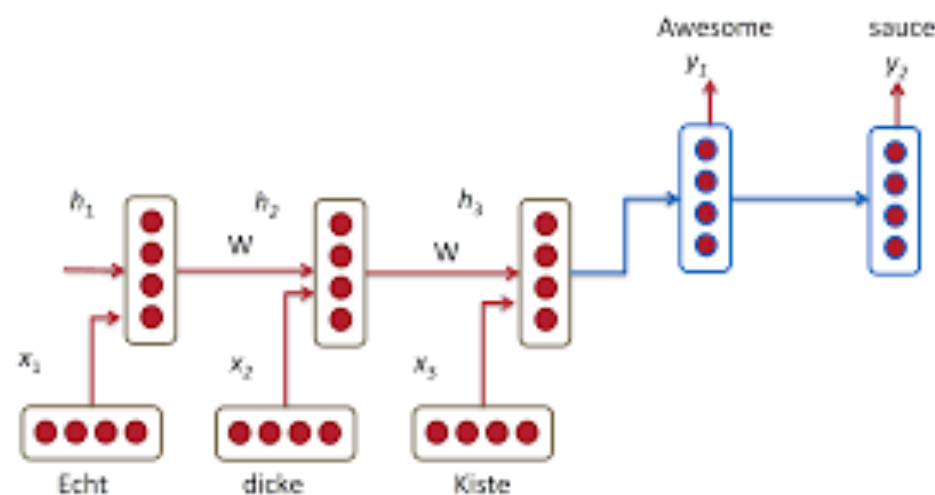
Источник: Goldberg, 2016

ЯЗЫКОВЫЕ МОДЕЛИ [LANGUAGE MODEL]

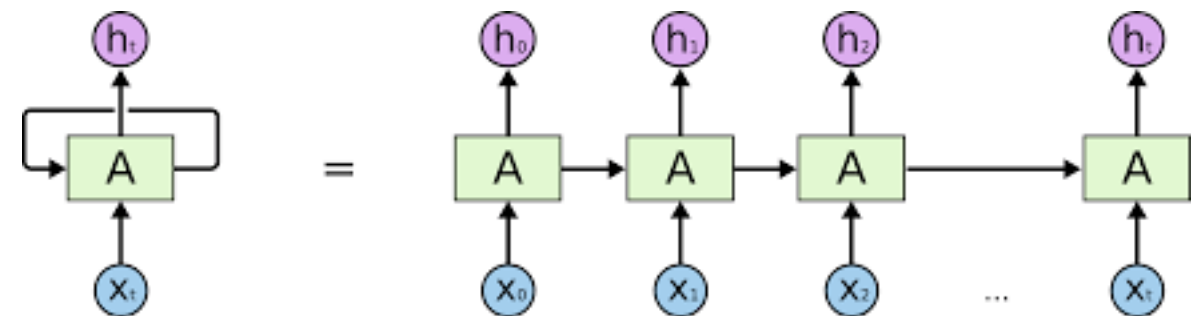
- Цепи Маркова
- Нейронные языковые модели
- Рекуррентные нейронные языковые модели
 - LSTM, GRU
- Seq2seq архитектуры



Источник: medium.com



Источник: wildml.com



Источник: [colah.github.com](https://colah.github.io)

КЛАССИФИКАЦИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ [SEQUENCE LABELLING]

- Условные случайные поля
- Рекуррентные нейронные сети

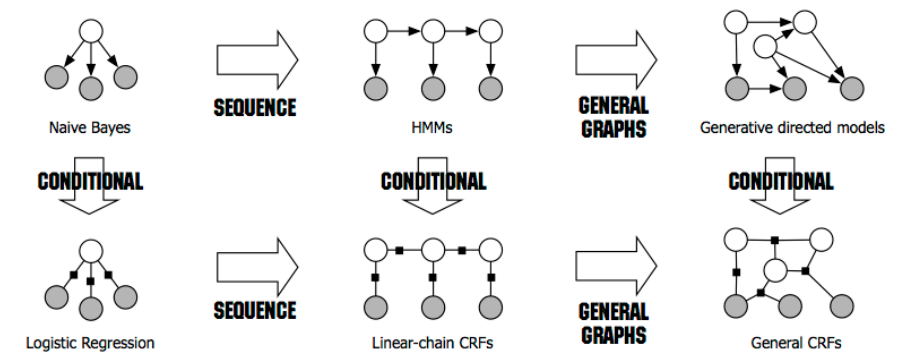


Figure 1.2 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs.

Источник: McCallum, 2012

Извлечение именованных сущностей

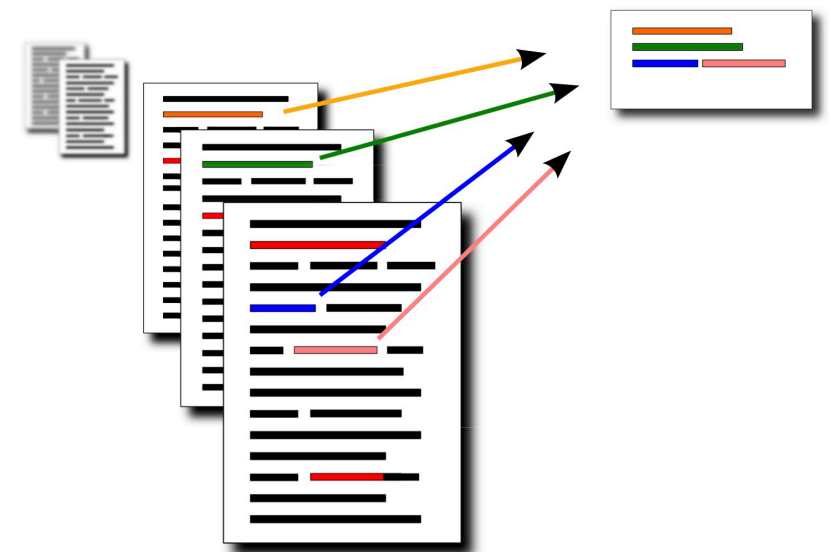


И МНОГОЕ ДРУГОЕ



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
- Томи́та-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.