

Семантический анализ текстовых документов

Денис Турдаков
turdakov@ispras.ru

План

- Зачем нужно анализировать текстовые документы?
- Какие возникают проблемы?
- Современные подходы к их решению.
- Решение одной из задач на примере системы Texterra

Классические проблемы

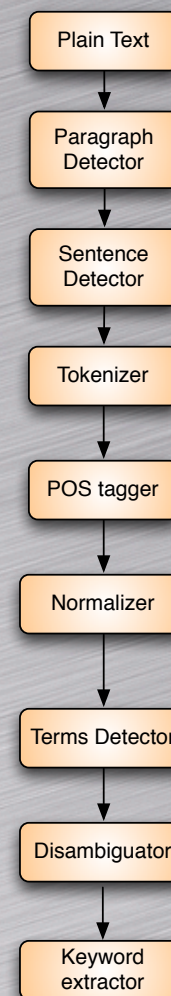
- Информационный поиск (IR)
- Извлечение информации (IE)
 - распознавание именованных сущностей (NERC)
- Вопросно-ответные системы (QA)
- Классификация и кластеризация
- Автоматическое аннотирование и реферирование
- Машинный перевод

Относительно новые задачи

- Информационный взрыв и масштабируемость
- Обработка полу-структурированных документов (таблиц, списков, ...)
- Обработка сообщений в социальных сетях
- Автоматическое построение онтологий
- Обогащение документов

Обработка текста

- Основная проблема - **многозначность**
- Морфология (определение части речи)
- Синтаксис
- Семантика
- Прагматика



Методы решения

- Описание знаний экспертами
- Методы основанные на лексико-семантических ресурсах
- Статистические методы.
Машинное обучение

Проект Texterra/Lognoo

TEXTERRA

- Система основана на новых методах обработки текстов, использующих структурную и текстовую информацию Википедии
- <http://modis.ispras.ru/texterra>

Решаемые задачи

- Вычисление семантической близости между концепциями Википедии
- Разрешение лексической многозначности
- Выделение основных тем и ключевых терминов из документов
- Поиск в коллекциях документов

Википедия

Moscow State University

From Wikipedia, the free encyclopedia

Coordinates: 55°42′10″N, 37°31′50″E

"MGU" redirects here. For other uses, see [MGU \(disambiguation\)](#).

M.V. Lomonosov Moscow State University (Russian: Московский государственный университет имени М.В.Ломоносова, *romanised:* *Moskovskiy gosudarstvennyy universitet imeni M. V. Lomonosova*, often abbreviated **МГУ**, **MSU**, **MGU**) is the largest university in Russia. Founded in 1755, it claims to be the oldest university in Russia. As of 2004, the university has some 4,000 staff teaching 31,000 students and 7,000 postgraduates. Its current rector is Viktor Sadovnichiy.

Contents [hide]

- History
- The campus
- Faculties
- Institutions and research centres
- Famous alumni and faculty
- Partner universities
- See also
- References
- External links

History

[edit]



This article or section's coverage of a controversial issue may be inaccurate or *unbalanced* in favor of certain viewpoints.
Please *improve the article* by adding information on neglected viewpoints,

M.V. Lomonosov Moscow State University

Московский государственный университет имени М.В.Ломоносова



Moscow State University's main building

Established:	25 January 1755
Rector:	Viktor Sadovnichiy
Faculty:	4,000
Staff:	15,000
Undergraduates:	40,000
Postgraduates:	7,000
Location:	Moscow, Russia

Семантическая близость

- **Определение.** Семантической близостью называется отображение $f : X \times X \rightarrow \mathcal{R}$, ставящее в соответствие паре объектов действительное число
- **Свойства**
 - $0 \leq f(x, y) \leq 1$
 - $f(x, y) = 1 \Leftrightarrow x = y$
 - $f(x, y) = f(y, x)$

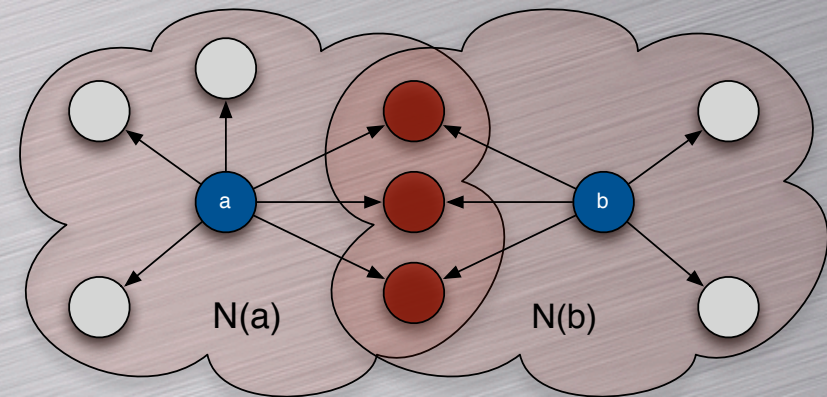
Локальные методы

- $$sim_{cos}(a, b) = \frac{|N(a) \cap N(b)|}{\sqrt{|N(a)|^2 + |N(b)|^2}}$$

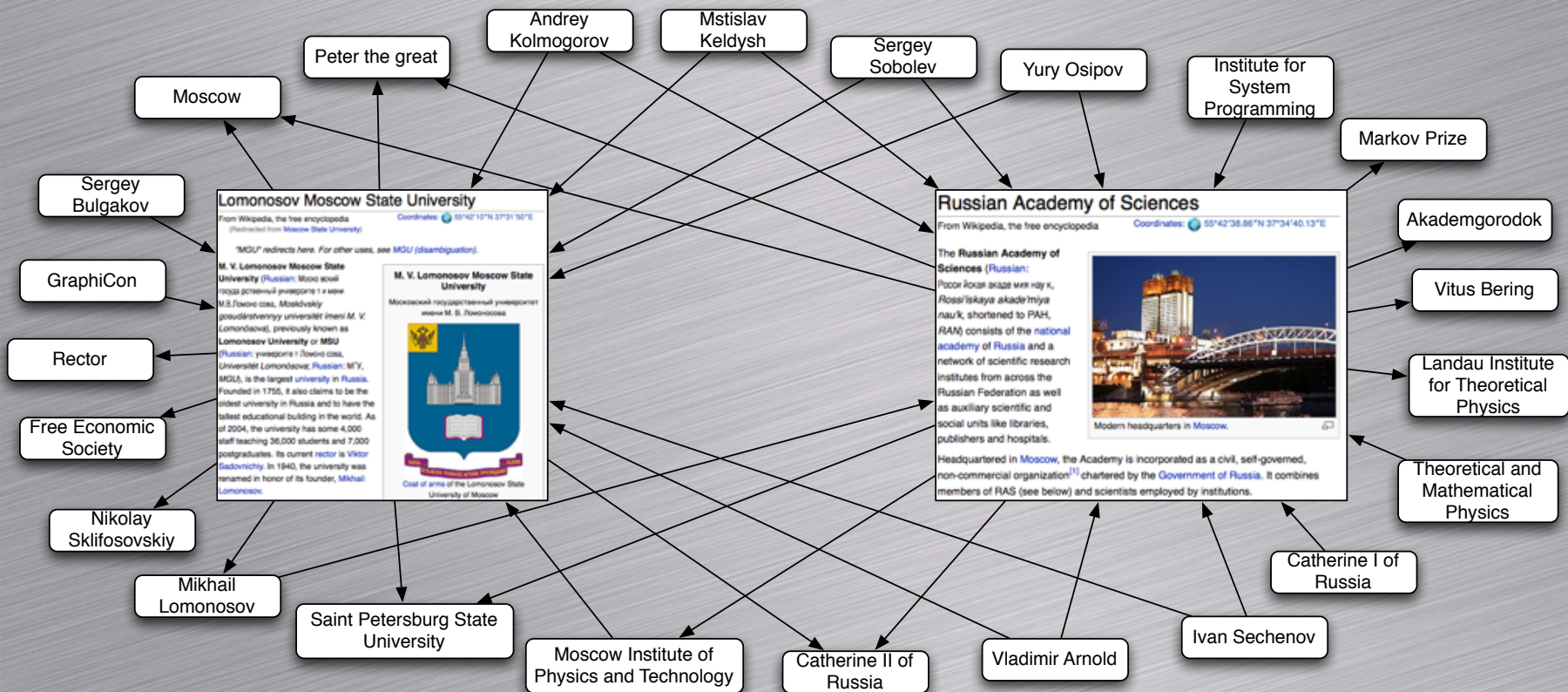
- $$sim_{Jaccard}(a, b) = \frac{|N(a) \cap N(b)|}{|N(a) \cup N(b)|}$$

- $$sim_{Dice}(a, b) = \frac{2|N(a) \cap N(b)|}{|N(a)| + |N(b)|}$$

- $$sim_{GD}(a, b) = \frac{\log(\max(|N(a)|, |N(b)|)) - \log(|N(a) \cap N(b)|)}{\log(|W|) - \log(\min(|N(a)|, |N(b)|))}$$



Пример



Глобальные методы

- Основаны на модели случайного блуждания (random walk)
- Методы адаптирующие PageRank для вычисления семантической близости
- SimRank (Jeh, Widom 2002)

Универсальная база знаний

- Словарь из терминов Википедии
- База данных возможных значений для многозначных терминов
- База данных возможных синонимов
- Взвешенный граф концепций с весами, отражающими семантическую близость между ними

[illegible]

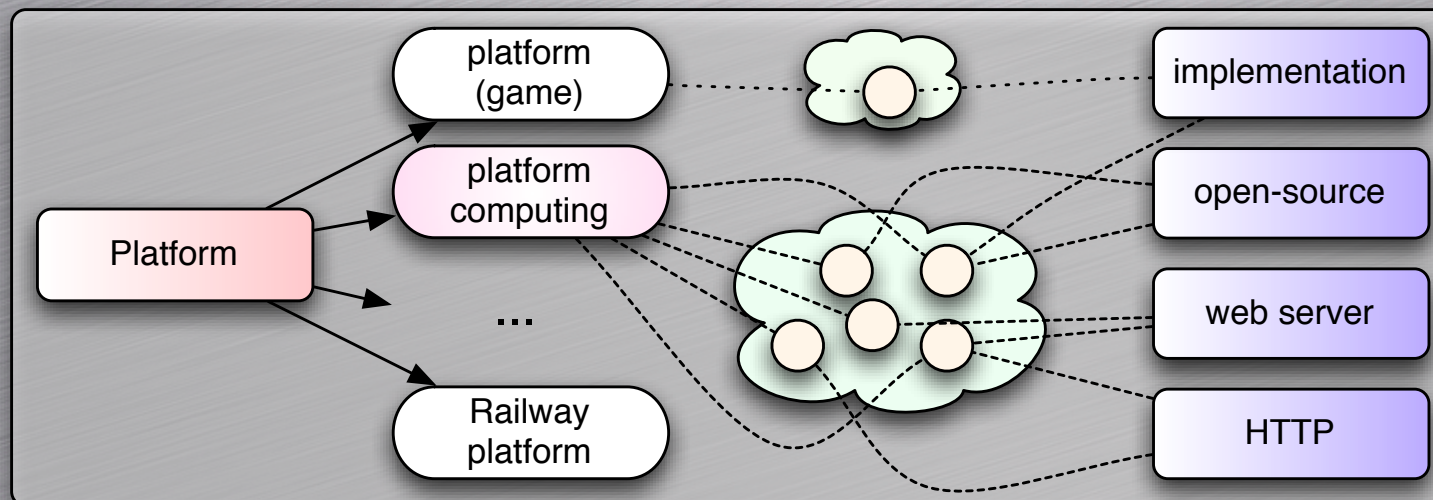
Устранение лексической многозначности

- Наиболее частое значение (MCS)
- Алгоритм Леска (1986): "PINE CONE"
 - PINE
 1. Kinds of **evergreen tree** with needle-shaped leaves
 2. Waste away through sorrow or illness
 - CONE
 1. Solid body which narrows to a point
 2. Something of this shape whether solid or hollow
 3. Fruit of certain **evergreen tree**
- $\text{PINE \#1} \cap \text{CONE \#3} = 2$

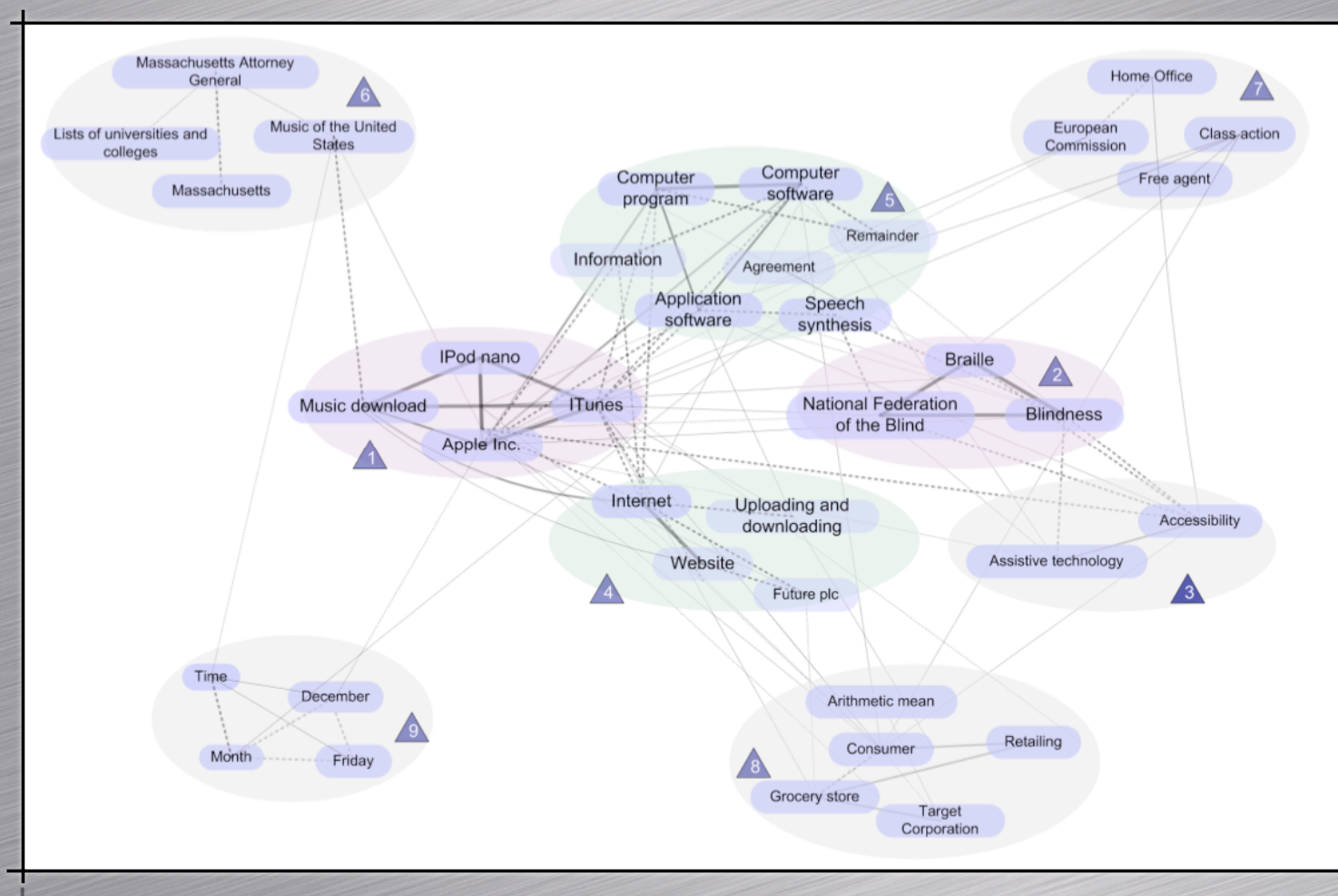
Метод, использующий однозначный контекст

D.Turdakov, P.Velikhov (SYRCoDIS 2008)

Jigsaw is W3C's **open-source** project that started in May 1996. It is a **web server platform** that provides a sample **HTTP 1.1 implementation** and ...

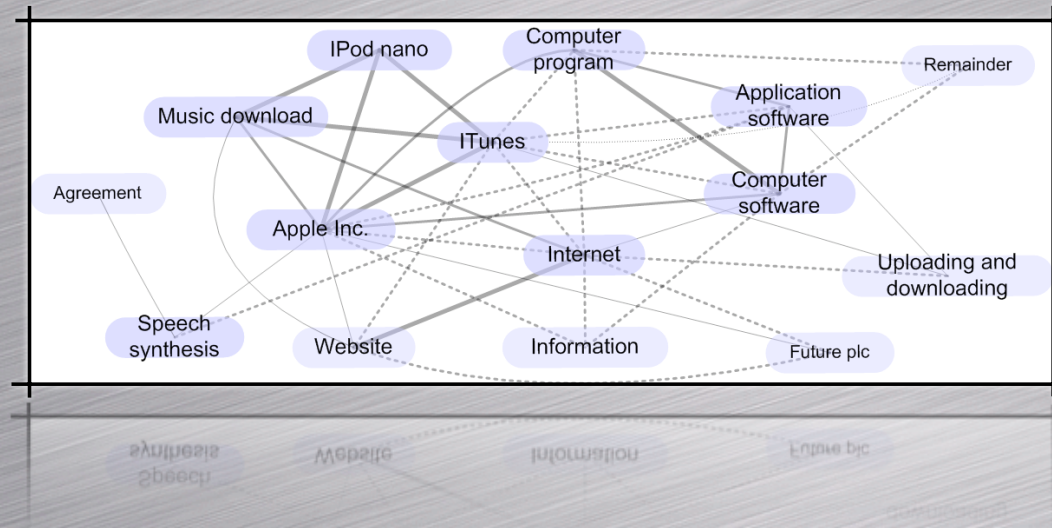


Выделение основных тем



Ранжирование терминов

- Задача: отсортировать термины в документе по их семантической важности
- Критерий важности: центральность вершины в графе терминов в документе



- Вес термина: алгоритм PageRank на взвешенном графе

$$V(a) = (1 - d) \times N_i + d \times \sum_{c_i \in N(a)} \frac{V(c_i) \times w_i}{|w_{c_i}|}$$

Поиск документов: Blognoon

- Использование словаря концепций Wikipedia как основы для формирования поисковых запросов
 - Статистика Google – 40% запросов ведут к статье Wikipedia как к первому результату
 - Соответствует “исследовательскому” поиску
- Необходимость комбинировать поиск по концепциям и по терминам
 - Не все концепции есть в Wikipedia
 - Поддержка “навигационного” поиска
- Синтаксис
 - Electric car; Speed; fastest
 - Electric car; “fastest speed”

Поисковые запросы

BlogNoon

Try for example: [Clean tech](#)

iPhone

iPhone
iPhone games
iPhone Operating System
iPhone OS (4.0)
iPhone OS 4
iPhoneOS
iPhone/iPod Touch App Store
iPhone and iPod touch homebrew installers
iPhone jailbreaking
iPhone OS Jailbreaking

Explore

Explore the BlogNoon

1. Enter your query

BlogNoon preferences
find your query

[Advanced Usage](#)

2. Navigate through the results

[Advanced Usage](#)

3. Click on a post title or blog name to read it and get recommendations

press Explore

Enter any keywords if you did not find any

queries

3. Click on a post title or blog name to read it and get recommendations

[Advanced Usage](#)

2. Navigate through the results

1. Enter your query

Фасетный интерфейс

BlogNoon

iPhone:

Explore

☒ Search posts ☐ Search blogs

Your query: [iPhone](#) ;

Sorted by [relevance](#) | [date](#)

iPad 2 Review

from [GottaBeMobile : Tablet PC, Netbook, Mobile Phones, and Multi-Touch News & Reviews](#) - 14.03.2011

intended for things like FaceTime or other video conferencing. It will be nice to snap photos of notes in a well lit meeting to import to Evernote. More on the iPad 2 camera quality further down the page. Another difference is the iPad 2 has a gyroscope just like **the iPhone 4**. The original iPad did have

[iPad](#) ; [iMovie](#) ; [App Store](#) ; [Wi-Fi](#) ; [Bluetooth](#) ; [iPhone](#) ; [iTunes](#) ;

iPad 2 Review

from [SlashGear - Feeding Your Gadget and Tech Obsessions](#) - 10.03.2011

place, but we're happy to be able to change it back. Of course, you won't need an iPad 2 to get a taste of iOS 4.3, and in fact you don't even need an iPad at all. Apple released the updated iOS earlier today for the GSM **iPhone 4**, **iPhone 3GS**, 3rd and 4th generation iPod touch and the original iPad

[iPad](#) ; [App Store](#) ; [iPhone](#) ; [Camera](#) ; [Wi-Fi](#) ; [iPod Touch](#) ; [Safari \(web browser\)](#) ;

MacWorld Top Apps

from [The Big Picture](#) - 03.03.2011

Best App Ever Angry Birds, published by Clickgamer.com Most Innovative App - **iPhone** Division Winner :: Word Lens by Quest Visual - iPad Division Winner :: Flipboard by Flipboard Inc. Best Visual Design - **iPhone** Division Winner :: Camera+ by tap tap tap - iPad Division Winner :: Flipboard by

[iPhone](#) ; [iPad](#) ; [Angry Birds](#) ; [Plants vs. Zombies](#) ; [Stitcher Radio](#) ; [Navigon](#) ; [Chair Entertainment](#) ;

How to Switch from AT&T to the Verizon iPhone 4

from [Media 3.0 with Shelly Palmer](#) - 28.02.2011

to a two-year contract for a 7-month old, 3G phone that is already obsolete. 4G phones are around the corner, and you can be sure Apple will be selling one this June – notice that **the iPhone 3GS** came out in June 2009 and **the iPhone 4** came out in June of 2010. Let's see ... hummm ... **iPhone 5** in June

[iPhone](#) ; [United States dollar](#) ; [EBay](#) ; [AT&T](#) ; [Verizon Communications](#) ; [Emmy Award](#) ; [NBC Universal](#) ;

Your Next Query

[relevance](#) | [bundles](#)

Mobile operating systems

[iPad](#)
[App Store](#)
[iPod Touch](#)
[iOS \(Apple\)](#)
[Angry Birds](#)

Cloud clients

[Android \(operating system\)](#)
[Smartphone](#)
[Nokia](#)
[Motorola](#)
[Best Buy](#)
[BlackBerry](#)
[Google Voice](#)
[Research In Motion](#)
[TFT LCD](#)
[Samsung Group](#)
[Opera Mini](#)

Windows media players

[iTunes](#)
[Twitter](#)
[Apple TV](#)
[iTunes Store](#)
[Netflix](#)
[Blog](#)
[Podcast](#)

[iPhone](#) ; [United States dollar](#) ; [EBay](#) ; [AT&T](#) ; [Verizon Communications](#) ; [Emmy Award](#) ; [NBC Universal](#) ;

[Podcast](#)
[Blog](#)

Рекомендации

BlogNoon

☒ Search posts ☐ Search blogs ☐ Search in blog **The Big Picture**

Explore

The Big Picture

This blog is about [All posts](#)

Berkshire Hathaway

[Warren Buffett](#) [Berkshire Hathaway](#)

IOS (Apple)

[iPhone](#) [iPad](#)

Real property law

[Negative equity](#) [Mortgage loan](#) [Debt-to-income ratio](#) [Information asymmetry](#)

CNBC

[United States](#) [Conspiracy theory](#) [Sony Pictures Classics](#) [Saudi Arabia](#) [Jim Cramer](#) [McKinsey & Company](#) [Wharton School of the University of Pennsylvania](#) [CNBC](#)

Pearl River Delta

[Hong Kong](#) [Hong Kong dollar](#) [Macau](#) [People's Republic of China](#)

Money

[Excess reserves](#) [Quantitative easing](#)

Economic problems

[Inflation](#) [Hyperinflation](#) [Core inflation](#) [NAIRU](#) [Output gap](#) [Deficit](#)

Paramount films

[The Weinstein Company](#) [Fox Searchlight Pictures](#) [Warner Bros.](#) [Paramount Pictures](#) [Coen brothers](#) [True Grit \(1969 film\)](#)

Monetary policy

[Federal Reserve System](#) [Central bank](#) [Monetary policy](#) [Great Depression](#) [Federal Open Market Committee](#)

Affordable housing

[Freddie Mac](#) [Fannie Mae](#)

Alternative investment management companies

[Bank of America](#) [Goldman Sachs](#) [Lehman Brothers](#) [William K. Black](#) [Accredited investor](#) [Hedge fund](#)

Recommended Blogs

GottaBeMobile : Tablet PC, Netbook, Mobile Phones, and Multi-Touch News & Reviews

IPad ; iPhone ; Steve Jobs ; Motorola ; Tablet personal computer ; Best Buy ; Android (operating system) ; Apple Store ; App Store ; Smartphone ;

SlashGear - Feeding Your Gadget and Tech Obsessions

IPad ; iPhone ; Nintendo 3DS ; App Store ; Android (operating system) ; iPod Touch ; Apple Store ; Intel Corporation ; Panasonic Corporation ; Motorola ;

TUAW -- The Unofficial Apple Weblog

IPad ; iPhone ; Mac OS X ; Apple Store ; App Store ; iPod Touch ; Steve Jobs ; iTunes ; MacBook Pro ; SK Telecom ;

Sizlopedia - Blog on Technology and New Media

IPad ; Steve Jobs ; Samsung Galaxy Tab ; iPhone ; Mac OS X ; MacBook Pro ; Best Buy ; MobileMe ; iPod Touch ; Apple Inc. ;

Примеры интересных проектов

- Qwiki.com
- Google translate
- Blognoon.com

Что дальше?

- Рассел, Норвиг. Искусственный интеллект. Современный подход.
- Jurafsky, Martin. Speech and Language Processing
- Barabasi. Linked ...
- Открытые API (Alchemy, OpenAmplify,..)
- Hadoop
- NLTK (Natural Language Tool Kit for python)
- <http://machinelearning.ru>
- <http://modis.ispras.ru/seminar>

Спасибо

Вопросы?

