

Основы обработки текстов

Лекция 9

Лексическая семантика

Возможные взгляды на семантику

- Лексическая семантика
 - значение индивидуальных слов
- Композиционная семантика
 - как значения комбинируются и определяют новые значения для словосочетаний
- Дискурс или прагматика
 - как значения комбинируются между собой и другими знаниями, чтобы задать значение текста или дискурс

План

- Основные понятия
 - слова и отношения между ними
 - словари и тезаурусы
- Вычислительная семантика
 - Разрешение лексической многозначности
 - Привязка к базам знаний

ОСНОВНЫЕ ПОНЯТИЯ

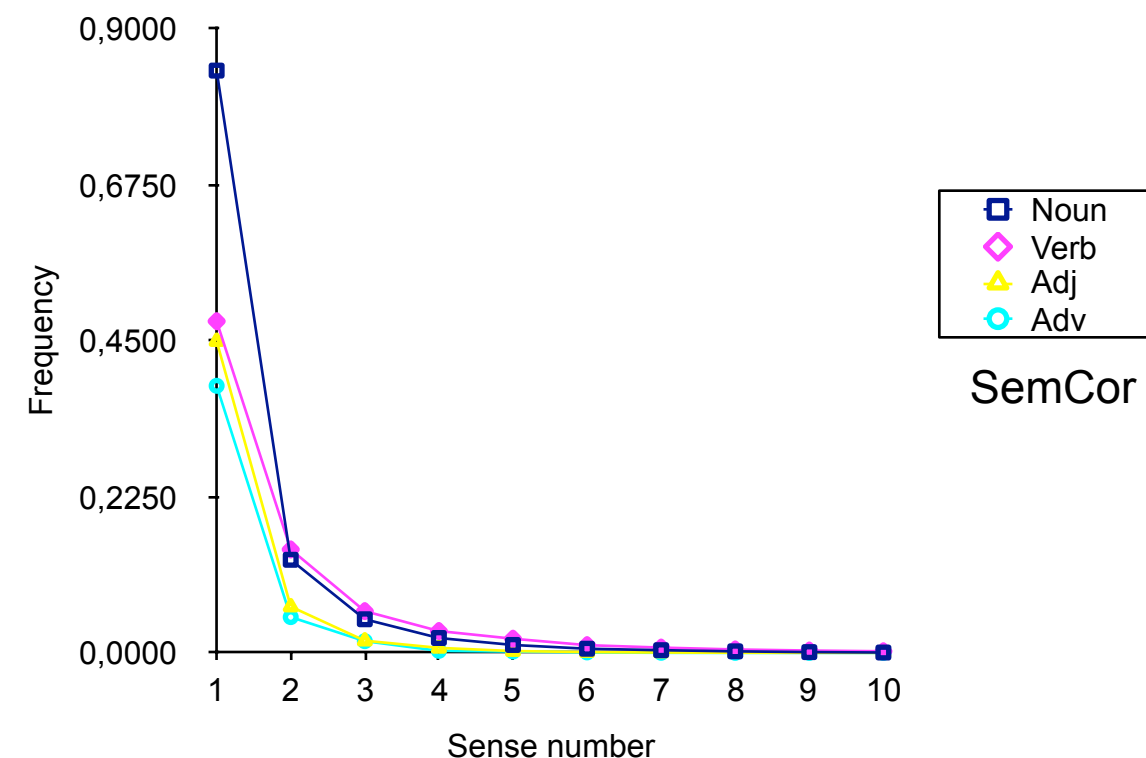
- Значение слова и многозначность
- Омонимия VS многозначность
 - ключ
 - платформа
- Метонимия
 - Я три *тарелки* съел
- Зевгма
 - За окном шел снег и рота красноармейцев
- Типы омонимов
 - омофоны (луг-лук, плод-плот)
 - омографы (м'ука - мук'а, гв'оздик-гвозд'ик)

Отношения между словами

- Синонимия
 - Машина / автомобиль
- Антонимия
 - большой / маленький, вверх / вниз, ложь / истина
- Обобщение и детализация (hyponym and hypernym/superordinate)
 - машина - транспортное средство
 - яблоко - фрукт
- Меронимы (партонимы) и холонимы
 - колесо - машина

Многозначность на практике

- Text-to-Speech
 - омографы
- Информационный поиск
- Извлечение информации
- Машинный перевод
- Закон Ципфа (Zipf law)



Основные вопросы

- Что такое значение?
 - Сколько значений у слова “платформа”?
- Что нужно для того, чтобы понять значение?
 - Что такое контекст?
 - Как определить связь значения с контекстом?
 - Разреженность языка

Базы знаний

- (В широком смысле) **База знаний** — база данных, содержащая правила вывода и информацию о человеческом опыте и знаниях в некоторой предметной области
- (В узком смысле) **База знаний** — некоторое структурированное описание предметной области

WordNet

- База лексических отношений
 - содержит иерархии
 - сочетает в себе тезаурус и словарь
 - доступен on-line
 - разрабатываются версии для языков кроме английского (в т.ч. для русского)

Категория	Уникальных форм
Существительные	117,097
Глаголы	11,488
Прилагательные	22,141
Наречия	4,601

- <http://wordnet.princeton.edu/>
- <http://wordnet.ru/>

Формат WordNet

The noun “bass” has 8 senses in WordNet.

1. bass¹ - (the lowest part of the musical range)
2. bass², bass part¹ - (the lowest part in polyphonic music)
3. bass³, basso¹ - (an adult male singer with the lowest voice)
4. sea bass¹, bass⁴ - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass¹, bass⁵ - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass⁶, bass voice¹, basso² - (the lowest adult male singing voice)
7. bass⁷ - (the member with the lowest range of a family of musical instruments)
8. bass⁸ - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective “bass” has 1 sense in WordNet.

1. bass¹, deep⁶ - (having or denoting a low vocal or instrumental range)
 *”a deep voice”; ”a bass voice is lower than a baritone voice”;
 ”a bass clarinet”*

WordNet: oтнoшeния мeжду слoвaми

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	<i>breakfast</i> ¹ → <i>meal</i> ¹
Hyponym	Subordinate	From concepts to subtypes	<i>meal</i> ¹ → <i>lunch</i> ¹
Member Meronym	Has-Member	From groups to their members	<i>faculty</i> ² → <i>professor</i> ¹
Has-Instance		From concepts to instances of the concept	<i>composer</i> ¹ → <i>Bach</i> ¹
Instance		From instances to their concepts	<i>Austen</i> ¹ → <i>author</i> ¹
Member Holonym	Member-Of	From members to their groups	<i>copilot</i> ¹ → <i>crew</i> ¹
Part Meronym	Has-Part	From wholes to parts	<i>table</i> ² → <i>leg</i> ³
Part Holonym	Part-Of	From parts to wholes	<i>course</i> ⁷ → <i>meal</i> ¹
Antonym		Opposites	<i>leader</i> ¹ → <i>follower</i> ¹

Relation	Definition	Example
Hypernym	From events to superordinate events	<i>fly</i> ⁹ → <i>travel</i> ⁹
Troponym	From a verb (event) to a specific manner elaboration of that verb	<i>walk</i> ¹ → <i>stroll</i> ¹
Entails	From verbs (events) to the verbs (events) they entail	<i>snore</i> ¹ → <i>sleep</i> ¹
Antonym	Opposites	<i>increase</i> ¹ ⇔ <i>decrease</i> ¹

Иерархии WordNet

Sense 3

bass, basso --

(an adult male singer with the lowest voice)

=> singer, vocalist, vocalizer, vocaliser

=> musician, instrumentalist, player

=> performer, performing artist

=> entertainer

=> person, individual, someone...

=> organism, being

=> living thing, animate thing,

=> whole, unit

=> object, physical object

=> physical entity

=> entity

=> causal agent, cause, causal agency

=> physical entity

=> entity

Sense 7

bass --

(the member with the lowest range of a family of musical instruments)

=> musical instrument, instrument

=> device

=> instrumentality, instrumentation

=> artifact, artefact

=> whole, unit

=> object, physical object

=> physical entity

=> entity

Как “зачение” определяется в WordNet

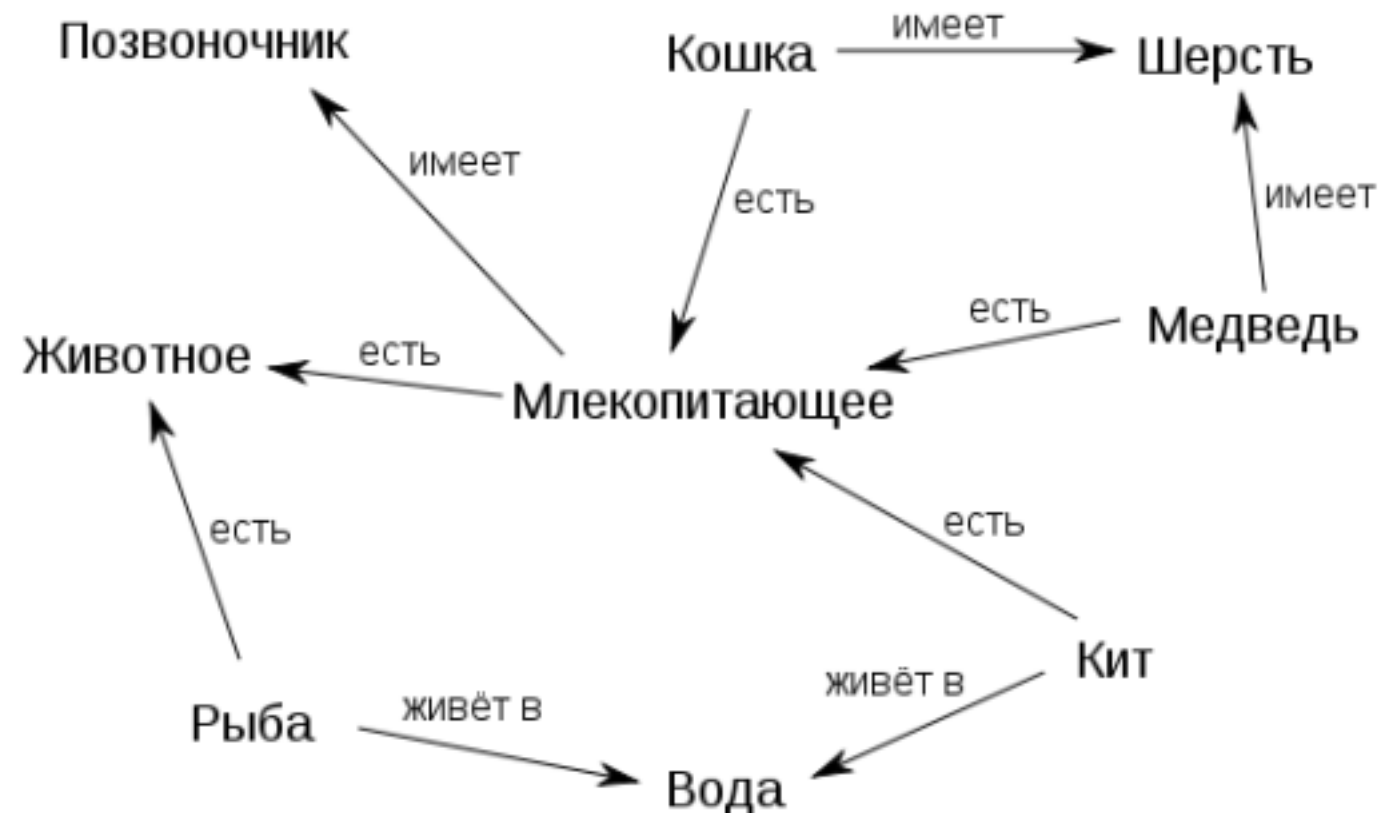
- Множественно синонимов называется **синсет**
- Пример

```
from nltk.corpus import wordnet
for synset in wordnet.synsets('chick'):
    print(synset.definition)
    print([lemma.name for lemma in synset.lemmas])
```

```
young bird especially of domestic fowl
['chick', 'biddy']
informal terms for a (young) woman
['dame', 'doll', 'wench', 'skirt', 'chick', 'bird']
```

Семантическая сеть

- **Семантическая сеть** — информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (рёбра) задают отношения между ними.



Викисловарь

- Викисловарь - <https://ru.wiktionary.org>.
- **Дом**
 1. архитектурное сооружение, предназначенное для жилья, и имеющее, как правило, стены, дверь и крышу
Просторный дом. Трёхэтажный дом.
 2. место, где кто-либо постоянно проживает
Здесь мой родной дом.
 3. офиц. совокупность жилых или производственных корпусов, а также служебных строений, расположенных на одном земельном участке и имеющих один учётный номер
 4. перен. фирма, предприятие
Торговый дом. Издательский дом.
 5. спорт. в кёрлинге — мишень в конце ледовой полосы, образованная четырьмя концентрическими кругами
 6. спорт. в бейсболе база, с которой начинается и которой заканчивается пробежка игрока
 7. перен. семья, династия, клан
Трёхсотлетие дома Романовых. Чума на оба ваши дома.

Викисловарь

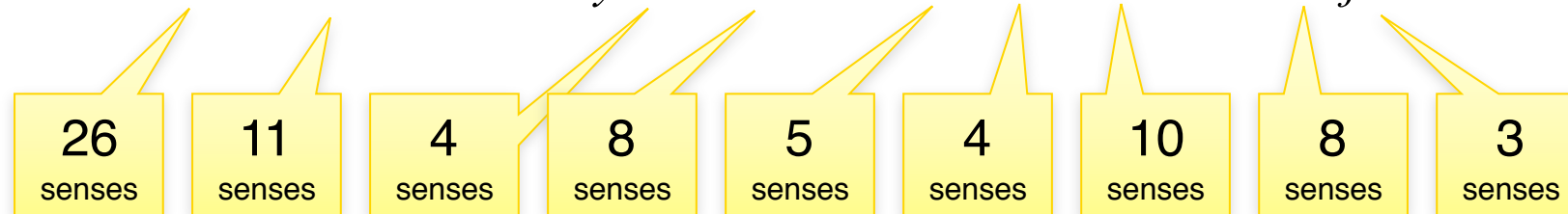
- Дом
 - Синонимы
 1. здание, корпус
 2. жилище, жильё, жилплощадь, резиденция
 3. —
 4. фирма, предприятие
 5. —
 6. —
 7. семья, династия, клан
 - Антонимы
 1. —
 2. —
 - Гиперонимы
 1. строение, постройка
 2. —
 3. учреждение
 4. —
 - Гипонимы
 1. барак, вилла, коттедж, многоэтажка, небоскрёб, хрущёба
 2. автодом, исправдом

Вычислительная лексическая семантика

- Разрешение лексической многозначности
- Привязка к базам знаний

Трудность разрешения лексической многозначности

I saw a man who is 98 years old and can still walk and tell jokes



43,929,600
senses

Разрешение лексической многозначности (РЛМ)

- Word Sense Disambiguation (WSD)
 - определение значения слова в контексте
 - обычно предполагается фиксированный список значений (например WordNet)
- Сводится к задаче классификации
- Отличается от задачи разграничения значений (word sense discrimination)

Разрешение лексической многозначности: варианты

- Определение значений только заранее выбранных слов (lexical sample task)
 - line - hard - serve; interest
 - Ранние работы
 - Обучение с учителем
- Определение значений всех слов (all-word task)
 - Проблема разреженности данных
 - Невозможно натренировать отдельный классификатор для каждого слова

Признаки

- Должны описывать **контекст**
- Предварительная обработка текста
 - параграфы, предложения, части речи, леммы, синтаксический разбор?
- Признаки в словосочетаниях с позициями
- Множества соседей
- Проблема разреженности языка
 - Использовать семантическую близость

Пример

*An electric guitar and **bass** player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.*

Collocational features	
word_L3	electric
POS_L3	JJ
word_L2	guitar
POS_L2	NN
word_L1	and
POS_L1	CC
word_R1	player
POS_R1	NN
word_R2	stand
POS_R2	VB
word_R3	off
POS_R3	RB

Bag-of-words features	
fishing	0
big	0
sound	0
player	1
fly	0
rod	0
pound	0
double	0
runs	0
playing	0
guitar	1
band	0

Алгоритмы

- Любые методы классификации

Вопрос на засыпку

- Как сделать классификатор для задачи определения значений всех слов (all-word task)?

Методы оценки

- Внешние (in vivo)
 - Машинный перевод с/без РЛМ
- Внутренние (in vitro)
 - Применение к размеченным данным (SemCor, SENSEVAL, SEMEVAL)
 - Измерение точности и полноты в сравнении со стандартными значениями
- Нижняя граница
 - Выбор случайных значений работает плохо
 - Более сильные границы: наиболее частое значение, алгоритм Леска
- Верхняя граница: согласие экспертов
 - 75-80 для задачи определения значений всех слов со значениями из WordNet
 - до 90% с менее гранулированными значениями

Методы основанные на словарях и тезаурусах

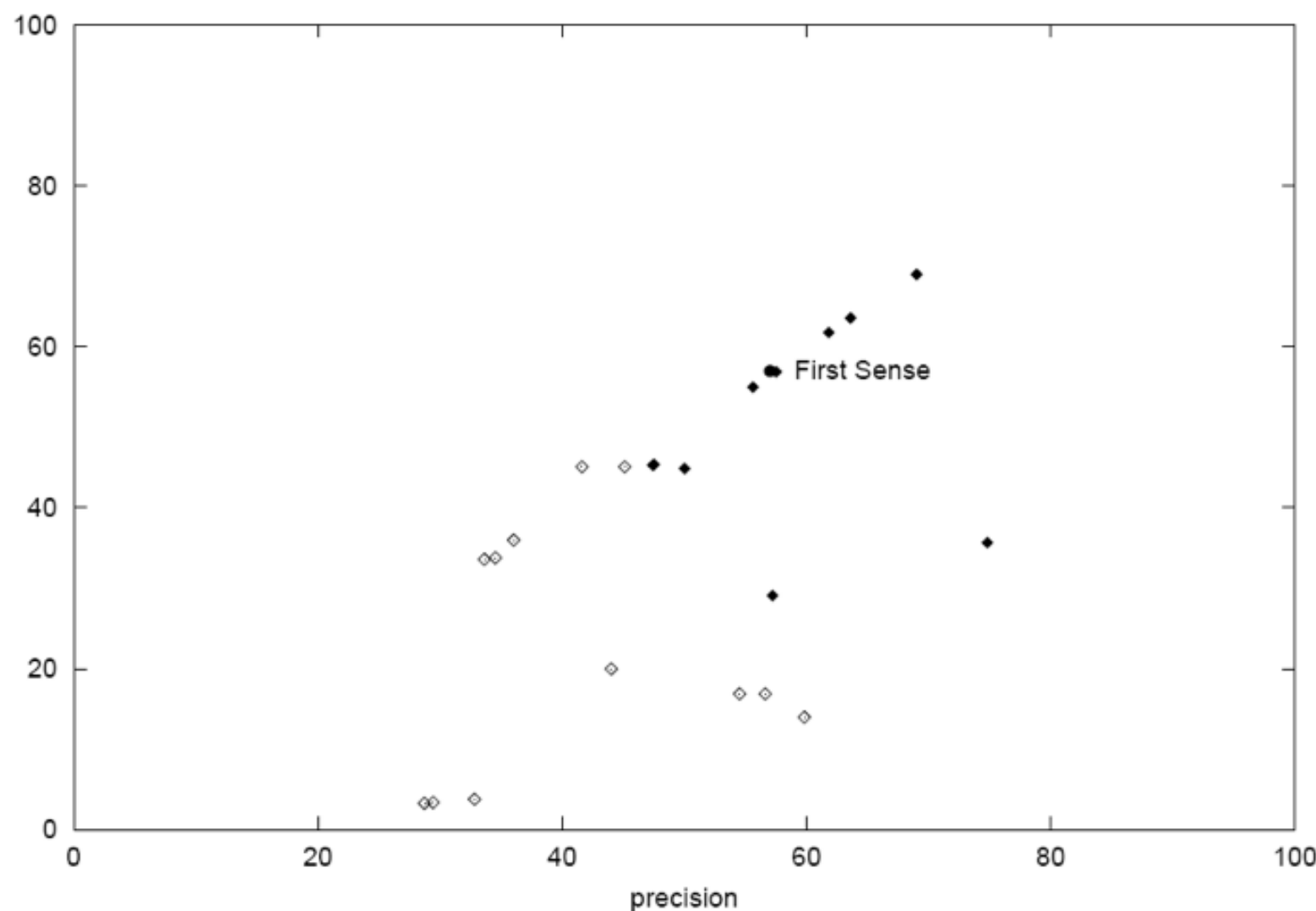
- Алгоритм Леска (1986)
 - Взять все определения целевого слова из словаря
 - Сравнить с определениями слов в контексте
 - Выбрать значение с максимальным пересечением
- Пример
 - pine
 1. a kind of **evergreen tree** with needle-shaped leaves
 2. to waste away through sorrow or illness
 - cone
 1. A solid body which narrows to a point
 2. Something of this shape, whether solid or hollow
 3. Fruit of certain **evergreen trees**
 - Определить значение: pine cone

Варианты алгоритма Леска

- Упрощенный (Simplified Lesk)
 - Взять все определения целевого слова из словаря
 - Сравнить со определениями словами в контексте
 - Выбрать значение с максимальным пересечением
- Корпусный (Corpus Lesk)
 - Включить предложения из размеченного корпуса в сигнатуру каждого значения
 - Взвесить слова через IDF
 - $IDF(w) = -\log P(w)$
 - Показывает лучшие результаты
 - Использовался как нижняя граница на SENSEVAL

Наиболее частое значение

- Сравнение методов на SENSEVAL-2



- McCarthy et. al. 2004 ACL - поиск наиболее частого значения по неразмеченному корпусу

Привязка к базам знаний

Связывание (именованных) сущностей

- Вход: неструктурированный текст
- Выход: Значение для каждой (именованной) сущности из базы знаний



База знаний. Википедия



Википедия
Свободная энциклопедия

[Заглавная страница](#)
[Рубрикация](#)
[Указатель А—Я](#)
[Избранные статьи](#)
[Случайная статья](#)
[Текущие события](#)

[Участие](#)
[Сообщить об ошибке](#)
[Сообщество](#)
[Форум](#)
[Свежие правки](#)
[Новые страницы](#)
[Справка](#)
[Пожертвовать](#)

[Инструменты](#)
[Ссылки сюда](#)
[Связанные правки](#)
[Служебные страницы](#)
[Постоянная ссылка](#)
[Сведения о странице](#)
[Цитировать страницу](#)


 Вы не представились системе
 [Обсуждение](#)
[Вклад](#)
[Создать учётную запись](#)
[Войти](#)

[Статья](#)
[Обсуждение](#)
[Читать](#)
[Текущая версия](#)
[Править](#)
[Править код](#)
[История](#)

Московский государственный университет

55°42′11″ с. ш. 37°31′50″ в. д. НГЯО

Материал из Википедии — свободной энциклопедии [\[править \]](#) [\[править код \]](#)

Текущая версия страницы пока не проверялась опытными участниками и может значительно отличаться от версии, проверенной 21 октября 2018; проверки требуют 7 правок.

Запрос «МГУ» перенаправляется сюда; см. также другие значения.

Моско́вский госуда́рственный университе́т имени М. В. Ломоно́сова — один из старейших^{[4][5]} и крупнейших^{[6][7]} классических университетов России, один из центров отечественной науки и культуры, расположенный в Москве.

С 1940 года носит имя **Михаила Васильевича Ломоносова**.

Полное наименование — Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М. В. Ломоносова». Широко используется аббревиатура «МГУ».

Университет включает в себя 15 научно-исследовательских институтов^[8], 43 факультета^{[9][10]}, более 300 кафедр и 6 филиалов (в их числе пять зарубежных — все в странах СНГ)^[11].

С 1992 года ректором МГУ является академик Виктор Антонович Садовничий.

Содержание [скрыть]

- История становления и развития Московского университета
 - Императорский Московский университет 1755—1917
 - Основание Московского университета в 1755 году
 - Московский университет в XVIII веке
 - Московский университет в XIX веке

Московский государственный университет имени М. В. Ломоносова
(МГУ имени М. В. Ломоносова)





Главное здание МГУ имени М. В. Ломоносова,
31 мая 2015 года

Значения

Основание Московского университета в 1755 году [править | править код]

Создание университета было предложено **И. И. Шуваловым** и **М. В. Ломоносовым**. Первоначально открытие университетских подготовительных работ, связанных в первую очередь с переводом, подписан императрицей **Елизаветой Петровной** 24 января 1755 года, утверждён проект об учреждении Московского Университета (XX—XXI веках — 25 января) отмечается как дата основания университета (25 января) 1755 года^[15]. Иван Иванович Шувалов — русский государственный деятель, генерал-адъютант (1760), фаворит императрицы Елизаветы I Петровны, меценат, основатель Московского университета и Академии художеств. Почётный директор. Ни в официальных документах, предположительно, не упоминается^[16]. По мнению историка **М. Т. Белявского** испортил ломоносовский проект, внося в него ряд изменений.

Ива́н Ива́нович Шувáлов — русский государственный деятель, генерал-адъютант (1760), фаворит императрицы Елизаветы I Петровны, меценат, основатель Московского университета и Академии художеств. Почётный директор.



Платформа

Материал из Википедии — свободной энциклопедии

[править | править код]

Текущая версия страницы пока не проверялась опытными участниками и может значительно отличаться от версии, проверенной 6 ноября 2016; проверки требуют 8 правок.

Платфо́рма (фр. *plateforme* от «плоский» + *forme* «форма») может означать:

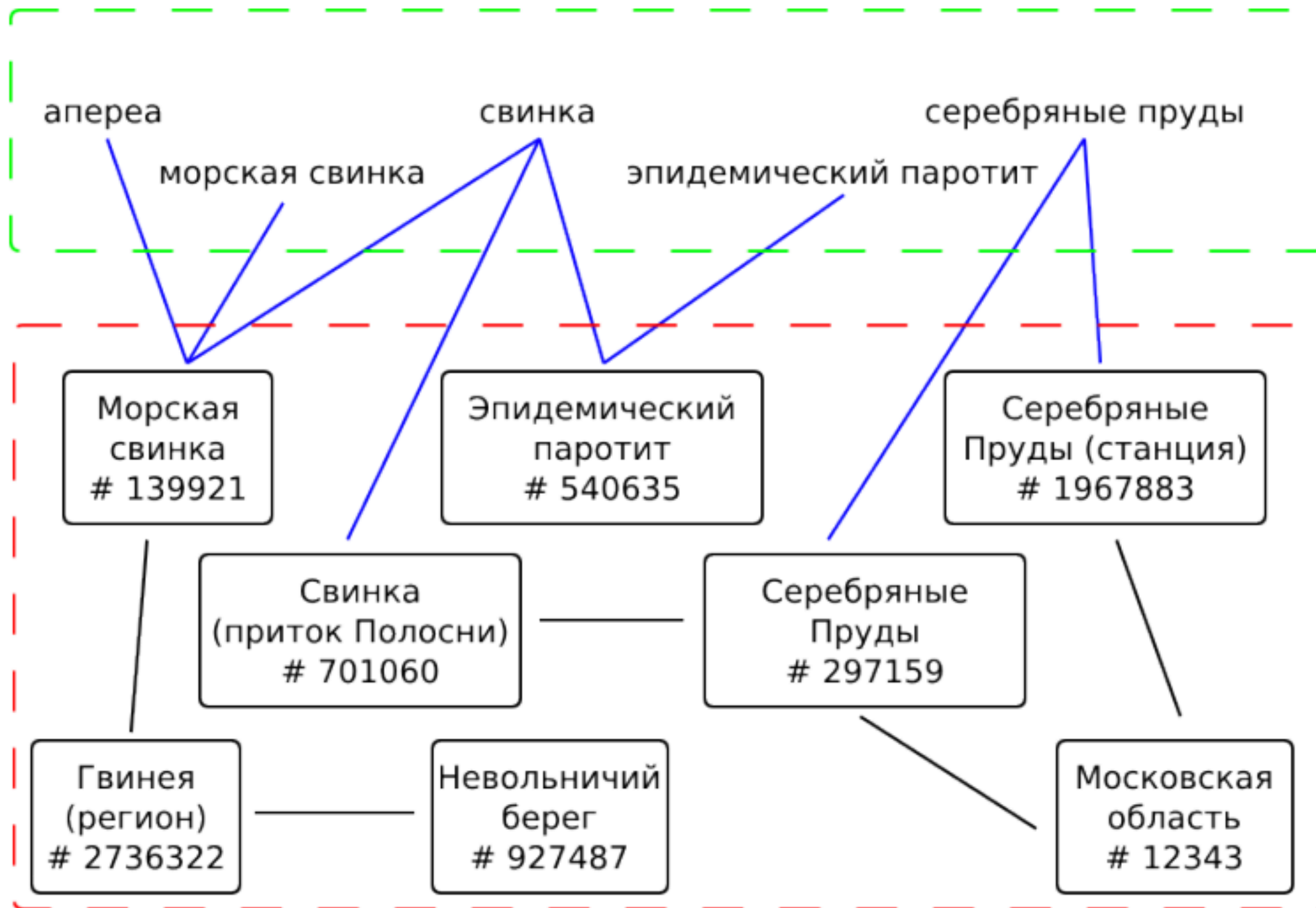
 Значения в Викисловаре

- Платформа** — возвышенная площадка, помост^[1] для посадки пассажиров в вагоны, для загрузки и выгрузки товаров.
 - Остановочный пункт** — предназначенный для остановки поездов участок железнодорожного пути, часто ошибочно метонимически называемый *платформой*
- Платформа** — в Англии и Америке публичная трибуна для произнесения речей, и в переносном смысле политическая речь — программа кандидата в депутаты пред выборами.
- Платформа** — совокупность основных компонентов, набор комплектующих, типовые конструктивные и технологические решения, применяемое оборудование в конструкции автомобиля.
- Платформа** — железнодорожный подвижной состав: род вагона, товарный вагон без крыши.
- Платформа** — крупный участок континентальной земной коры, характеризующийся относительно спокойным тектоническим режимом.
- Платформа** — один из компонентов *обуви*: заметно утолщённая *подошва обуви*.

Название [править | править код]

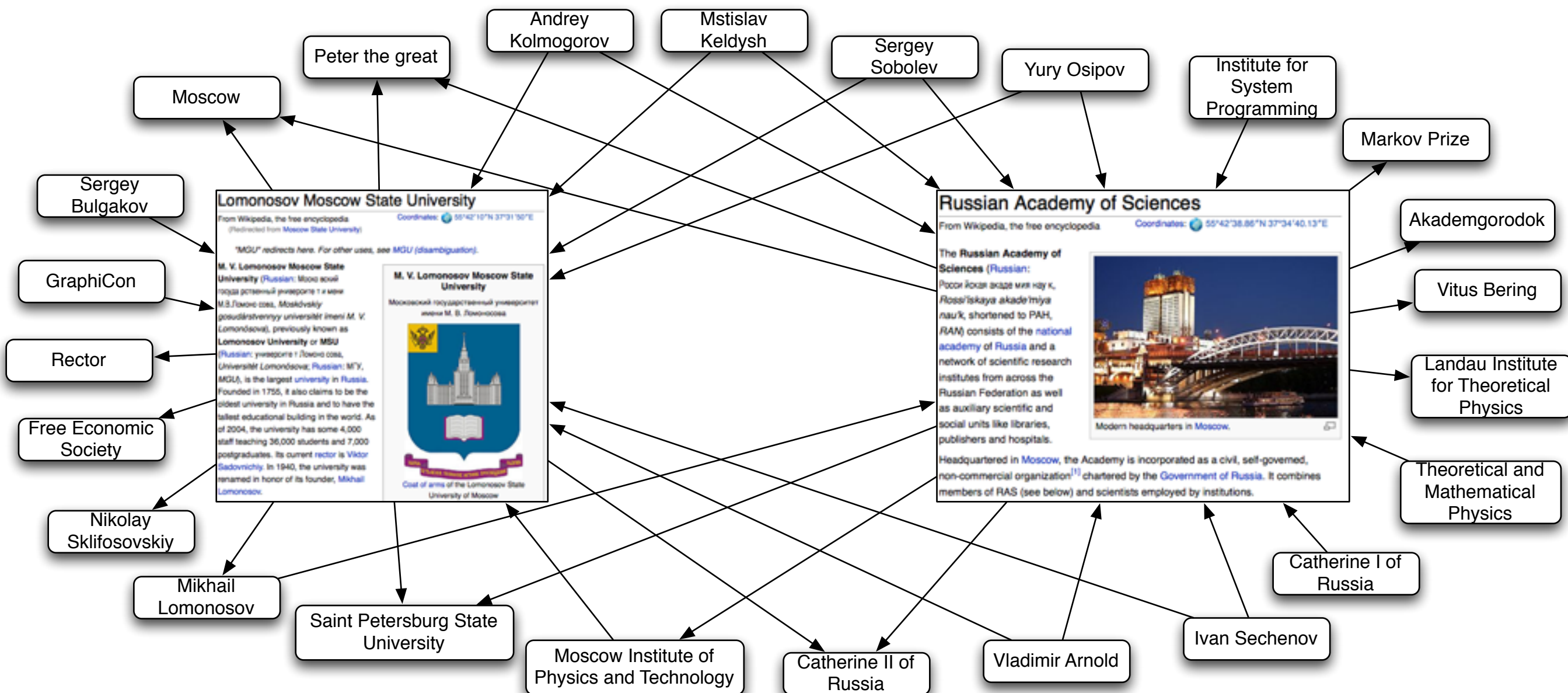
- «Платформа» — музыкально-литературный клуб в Санкт-Петербурге.
- «Платформа» — популярная американская стрижка.
- «Платформа» — роман французского писателя Мишеля Уэльбека, опубликованный в 2001 году.

База знаний. Википедия



Использование Википедии: семантическая близость

- Нормализованное количество общих соседей



- Близкие концепты чаще встречаются вместе

TEXTERRA

Названа **максимальная стоимость** проезда по трассе **Москва — Казань**.

Ведомство предложило построить трассу не до **Казани**, а до **Владимира**, и вместе с этим реконструировать существующую **трассу М7 Москва - Уфа**. Об этом сообщала газета "**Ведомости**" со ссылкой на правительственные источники.

Яндекс Новости # 01.11.2019 18:20

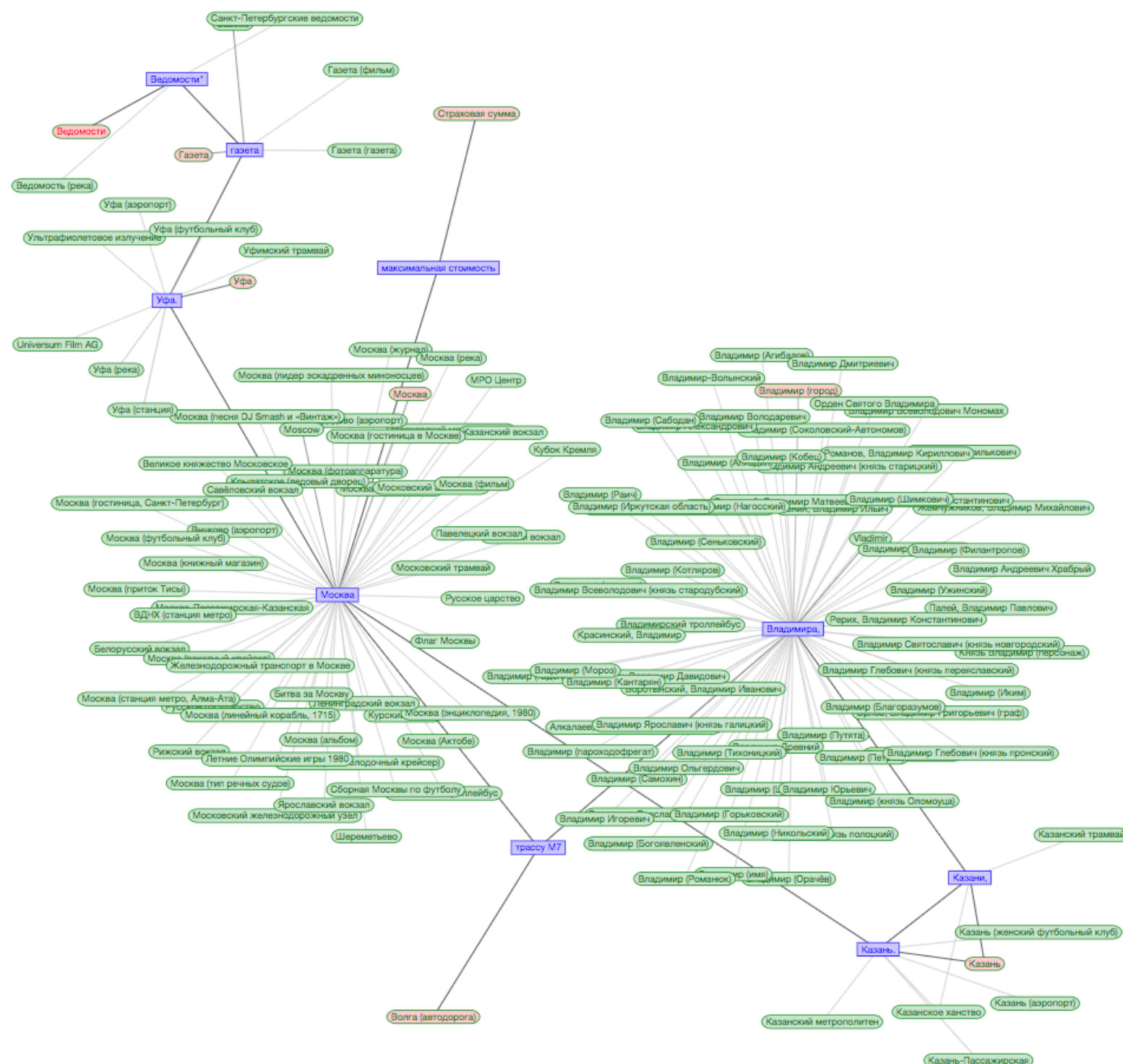
Найденные сущности

- **Ведомости**
- Газета
- Уфа
- Москва
- **Волга (автодорога)**
- Владимир (город)
- Казань
- Страховая сумма

Wikipedia

Федеральная автомобильная дорога М7 «Волга» — автомобильная дорога федерального значения Москва — Владимир — Нижний Новгород — Казань — Уфа (с подъездами к Владимиру, Иванову, Чебоксарам, Ижевску, Перми). Часть европейского маршрута **Е 22**, часть трассы от

VisOntia



<https://api.ispras.ru/demo/vizontia>

Основные проблемы

- Синонимия (вариации имен): New York, NY, Big Apple
- **Многозначность**: NY ? New York / New Year
- Отсутствие значения
- Скорость обработки
- Постоянное изменение базы знаний
- Много языков

Алгоритм Milne-Witten

- Иcпoльзует oднoзнaчнoе тeрминoв в кaчecтвe кoнтeкcтa

Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (DFS) is an **algorithm** for traversing or searching a **tree** **tree structure** or **graph**. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before **backtracking**.

Formally, DFS is an **uninformed search** that progresses by expanding the first child node of the search **tree** that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search **backtracks**, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a **LIFO stack** for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	2.57%	63.26%
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
...		

Milne, David, and Ian H. Witten. "Learning to link with wikipedia." *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008.

Алгоритм Milne-Witten

- Commonness (популярность значения)

$$commonness(e, m) = \frac{count(e, m)}{count(e)}$$

- Relatedness (расстояние до контекста)

$$relatedness(a, b) = \frac{\log \max(|A|, |B|) - \log |A \cap B|}{\log |W| - \log \min(|A|, |B|)}$$

a, b - концепты;

A, B - множества концептов, ссылающихся на e и s соответственно;

W - множество всех концептов.

- Вероятность термина быть ссылкой

$$link_prob(t) = \frac{count(t \text{ is link})}{count(t)}$$

Алгоритм Milne-Witten

- Семантическое расстояние от концепта a до контекста S :

$$distance(a, S) = \frac{\sum_{s \in S} w_s \times relatedness(a, s)}{\sum_{s \in S} w_s}$$

- Вес w_s концепта s :

$$w_s = \frac{1}{2} (link_prob(term(s))) + \frac{1}{|S|} \sum_{c \in S} relatedness(s, c)$$

- Качество контекста S :

$$quality(S) = \sum_{s \in S} w_s$$

Алгоритм Milne-Witten

- Признаки:
 - популярность значения - $\text{commonness}(e, m)$;
 - расстояние до контекста - $\text{distance}(e, S)$;
 - когерентность контекста - $\text{quality}(S)$.
- Алгоритмы машинного обучения:
 - Naïve Bayes;
 - дерево решений C4.5;
 - метод опорных векторов;
 - дерево решений C4.5 с бэггингом.
- Обучение / применение:
 - положительный пример — правильное значение термина;
 - отрицательные примеры — все остальные возможные значения термина;
 - на этапе применения выбирается концепт, с максимальной уверенностью классифицированный как правильный.

Понимание слов предложения ≠ понимание предложения

- Во время интервью:
 - Ваша главная слабость?
 - Правильно интерпретирую семантику вопроса, но игнорирую его суть.
 - Не могли бы вы привести пример?
 - Мог бы.

Что не было рассказано

- Композиционная семантика
- Представление знаний
- Семантические поля и семантические роли
 - PropBank
 - FrameNet
- Задача разграничения значений
- Автоматическое извлечение отношений между словами
- ...

Следующая лекция

- Языковые модели на основе нейронных сетей