

Анализ текста

Машинное обучение, 2017

Спасибо естественному свету разума.

Малютин Е. А.

Планчик

- предобработка текста
- представление
- некоторые задачи
- некоторые приложения

Предобработка текста

Зачем?

- убираем ненужное
- убираем шума
- снижаем устойчивость словаря
- помогаем алгоритмам ML
- выделяем "важное"

КАК?

- фильтрация (телефоны, email, html, etc)
- детект языка
- стемминг
- лемматизация
- стоп-слова
- дедупликация

Language Detecion

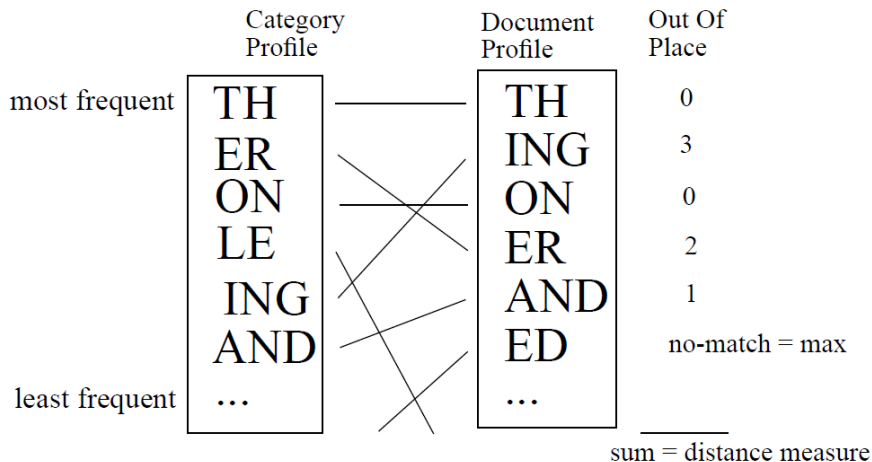


Рис.: Определение языка с помощью top-trigrams, см. langdetect (shyo, optimaize)

Стемминг – процесс нахождения основы слова. Основа слова – не обязательно морфологический корень. (Падающие => пада)

Лемматизация – нахождение нормальной, “канонической” формы слова. (Падающий => падать). По сравнению со стеммингом - лемматизация более “дорогая”, но “тонкая” операция.

Стемминг – простые rule-based алгоритмы (алгоритм Портера, Snowball stemmer)

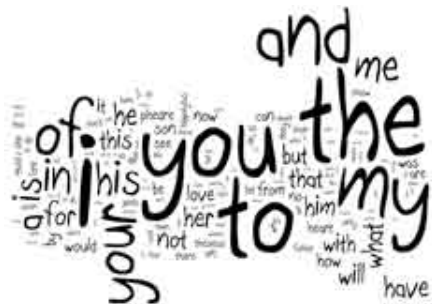
Лемматизация – словари, правила, филологи.

С помощью чего? nltk, pymorphy2

Stop-words

Зачем?

- **частотная гипотеза** – наиболее частые слова несут меньше всего информации
- составленные списки, срез по "частоте"
- минусы? есть



Дедупликация

Общая суть

- данные делятся на участки;
- выполняется поиск одинаковых участков (копий, дублей, повторов);
- все одинаковые участки, кроме первого, заменяются ссылками на первый участок.

А как?

- kd-деревья
- locality sensitive hashing
- autoencoders
- etc...

Векторизация

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

Рис.: Bag of words model

Минусы

- Теряется порядок слов
- Разреженные вектора
- Хранение словаря (или hashing trick)

TF-IDF model

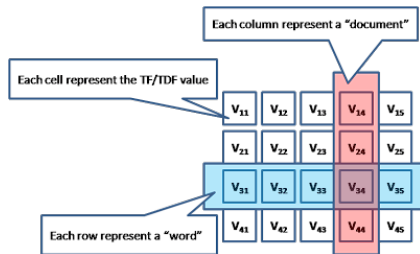
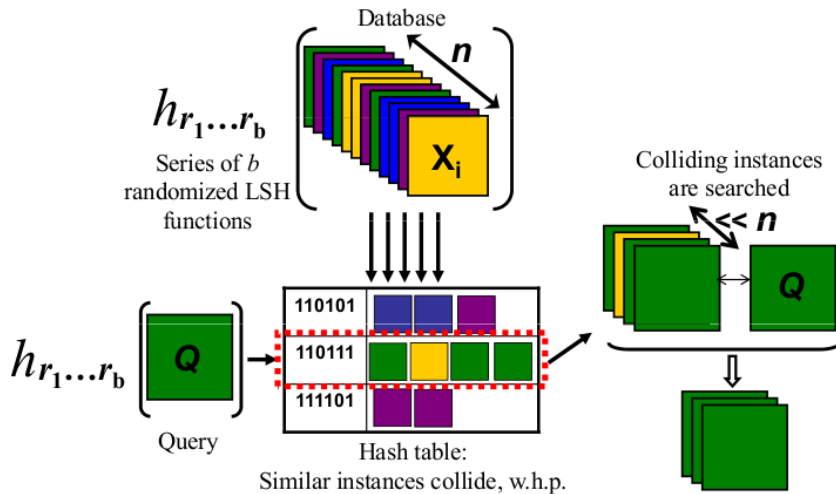


Рис.: TF-IDF text's representation

- $tf(t, d) = \frac{n_i}{\sum_k n_k}$
- $idf(t, D) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|}$
- уменьшает вес широкоупотребительных слов

Random Binary Projection

Обратно к дедупликации



Random Binary Projection

Обратно к дедупликации

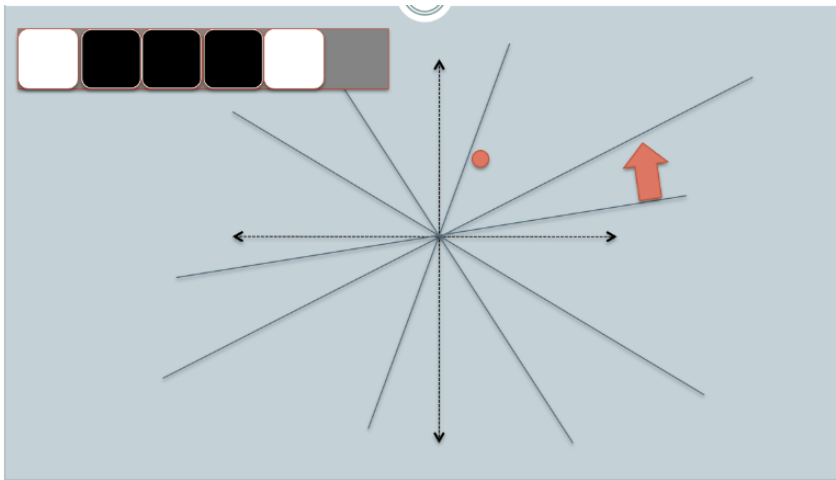


Рис.: Вид hash-функции

Word2Vec

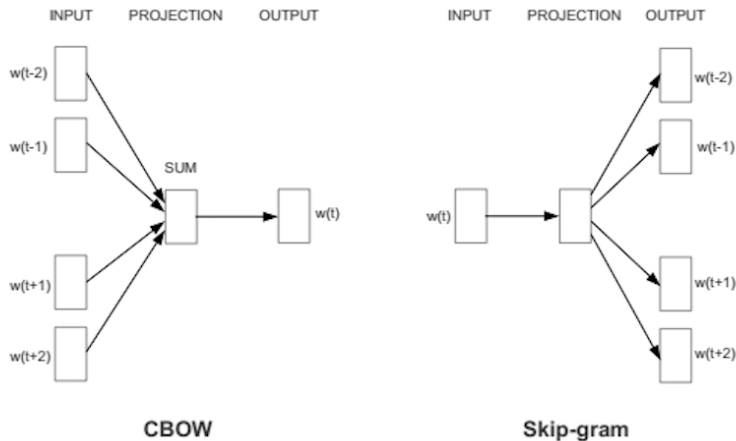


Рис.: Синтетическая задача

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

Рис.: Пример вывода, похожие слова по Mikolov

Word2Vec

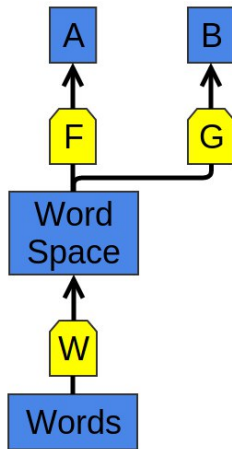


Рис.: Как с этим жить

Word2Vec

Пример в визуализации

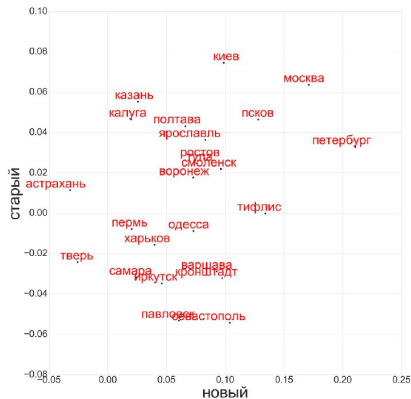


Рис.: НКРЯ 1897-1916

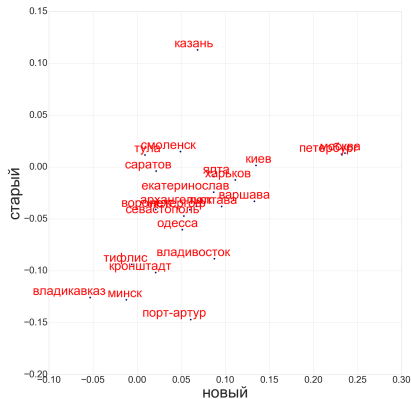


Рис.: НКРЯ 1917 -1929

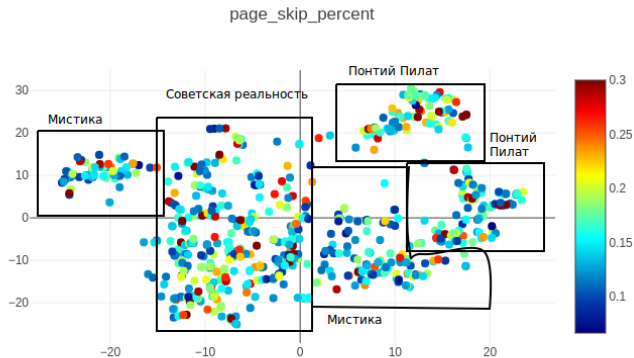


Рис.: Раскладка Мастера и Маргариты по word2vec с помощью t-sne

Где почитать?

- <https://habrahabr.ru/post/253227/> – совсем просто
- <https://www.kaggle.com/c/word2vec-nlp-tutorial/data> – сложнее
- <https://groups.google.com/forum/#!forum/word2vec-toolkit> – ещё сложнее

Анализ стиля (stylometry)

Приложения

- характеристика стиля
- кто написал это произведение?
- обнаружение плагиата
- ...

Идея

индивидуальные неконтролируемые признаки

Анализ стиля (stylometry)

Признаки

Lexical	Token-based (word length, sentence length, etc.)	Syntactic	Part-of-Speech
	Vocabulary richness		Chunks
	Word frequencies		Sentence and phrase structure
	Word n -grams		Rewrite rules frequencies
	Errors		Errors
Character	Character types (letters, digits, etc.)	Semantic	Synonyms
	Character n -grams (fixed-length)		Semantic dependencies
	Character n -grams (variable-length)		Functional
	Compression methods		

Рис.: Признаки в стилометрии, Stamatos, 2009

Compression based approach

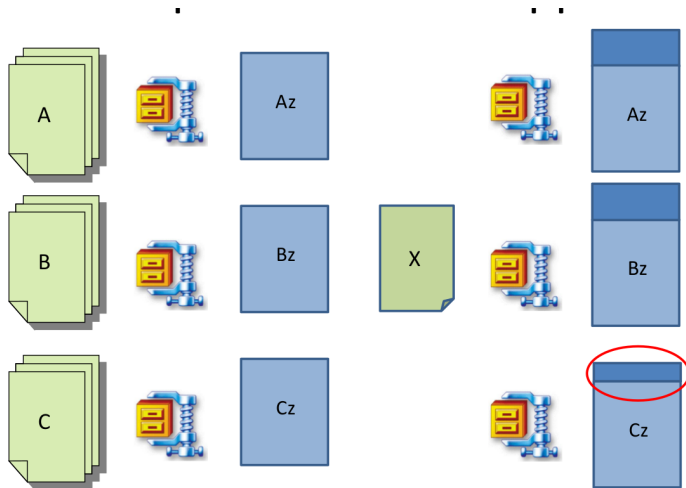


Рис.: Классификация на основе частотности, Дмитрий Хмельнёв

Красивые картинки

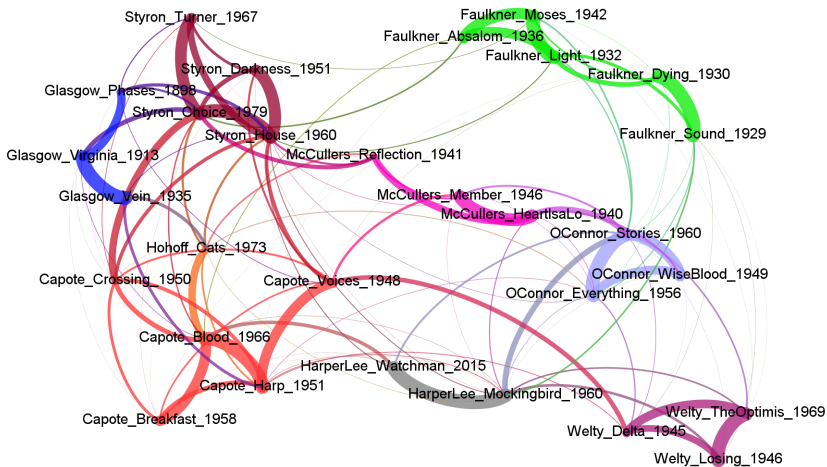


Рис.: Stylo, Jan Rybicki

Близкие задачи

- жанровый анализ
- характеристики автора (author profiling) – пол, возраст, родной язык
- удобочитаемость (readability)

Определение тональности

- определить тональность текста/предложения/твита
- +/0/-; шкала
- объект/субъект
- субъективность/тональность (отношение)/эмоции
(разочарование/грусть/...)/личностные характеристики/...

Методы

- классификация
- словари
- нейронные сети

Matthew Jockers, tonality plot

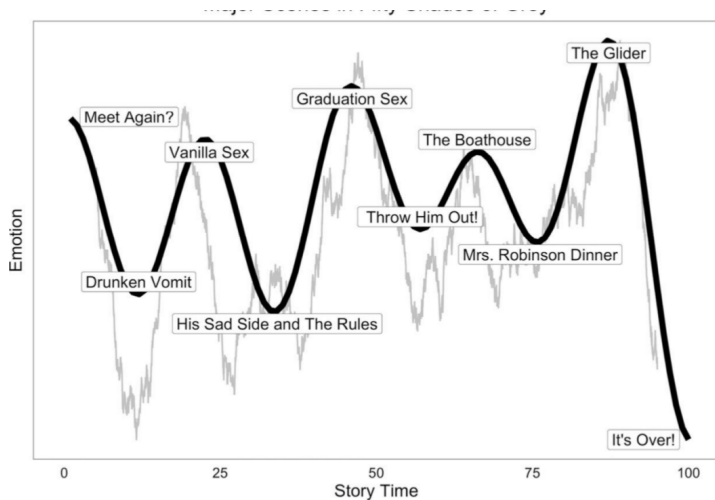


Рис.: 50 оттенков тональности

Matthew Jockers, tonality plot

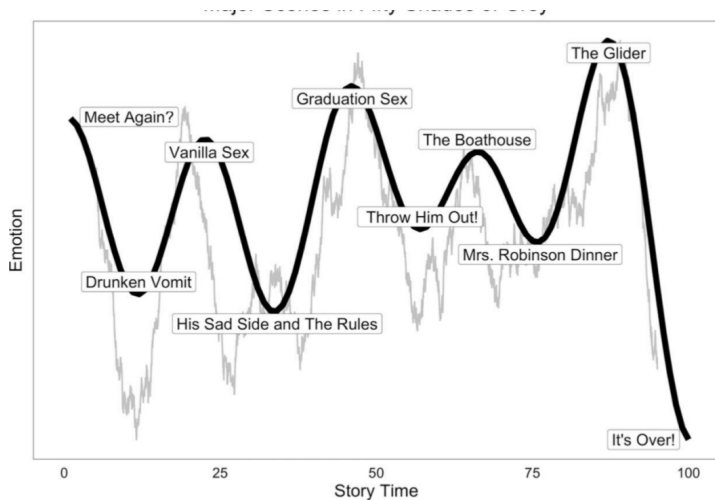


Рис.: 50 оттенков тональности

Matthew Jockers, tonality plot

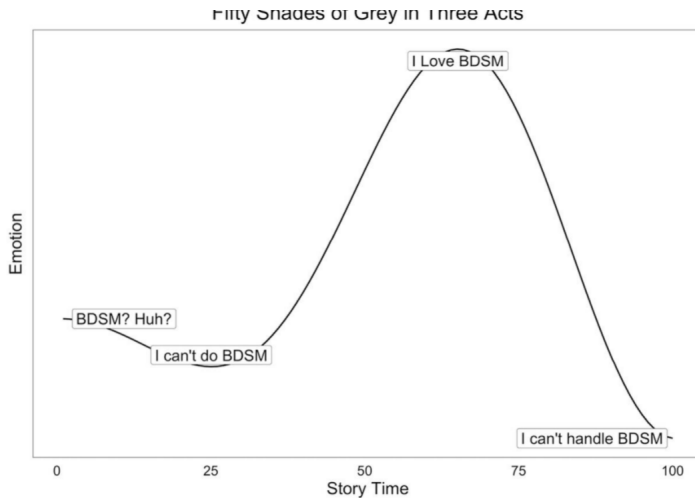
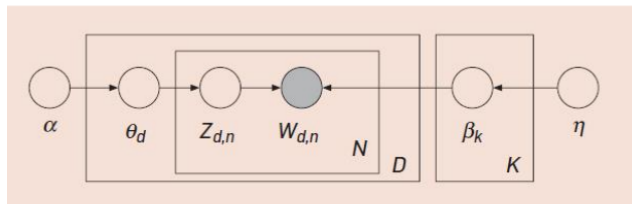


Рис.: 50 оттенков тональности в трёх актах

Латентное размещение Дирихле

Процесс порождения текста:

- 1 Случайным образом выбираем распределение тем (распределение Дирихле)
- 2 Для каждого слова в документе
 - Случайным образом выбираем тему из распределения из шага 1
 - Случайным образом выбираем слово из распределения слов в теме.



Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

Seeking Life's Bare (Genetic) Necessities

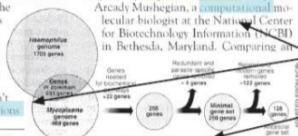
COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses 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**

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

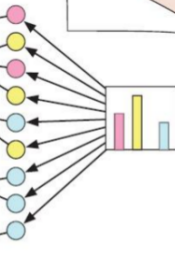
SCIENCE • VOL. 272 • 24 MAY 1996

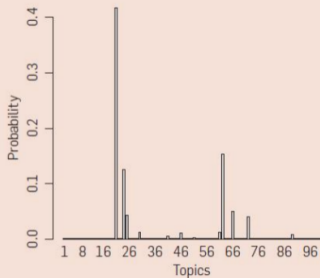
"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson at Uppsala University in Sweden. "We arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **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.

Topic proportions and assignments



**"Genetics"**

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

"Evolution"

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

"Disease"

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis

"Computers"

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

Применение

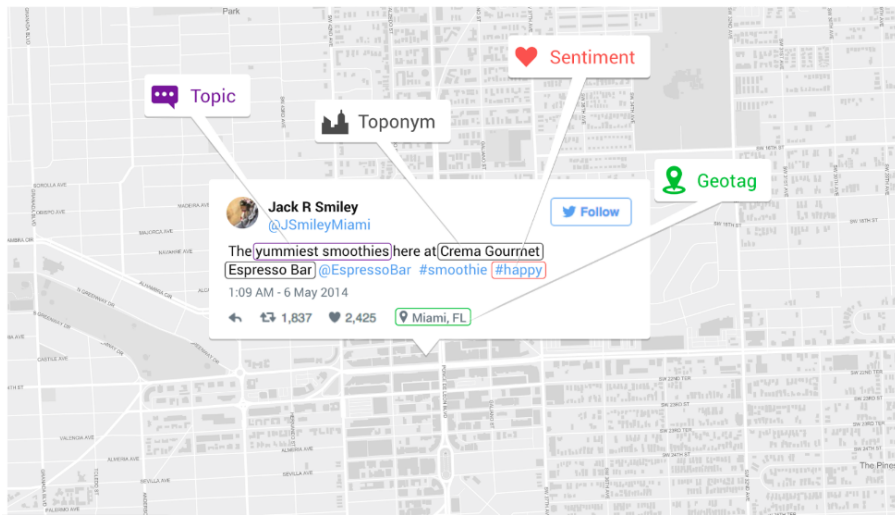


Рис.: Habidatum, урбанистика

Примение

MUSIC

WORK


INTERNET


FOOD


ENTERTAINMENT

WATCHING SPORTS

EVENTS

 enjoying some amazing #macarons at #janetteandco #nutella #venezuelanchocolate

 at burgerfi for awesome burgers! #burgerfi

 can't wait to dig into this yummy food @carnaval_miami !

MUSIC

WORK


INTERNET


FOOD


ENTERTAINMENT

WATCHING SPORTS

EVENTS

 @arionnation he has been our best bench guy this year (welp), but he has a lot of passed balls, which piss people off. also, he isn't gattis

 anybody who has a choice this week and chooses chip/joe over vin scully to watch braves at dodgers is doing baseball all wrong

 @mikenewmanns lol, the best humor comes from pain (fortunately here, that is just baseball fan pain, not real life pain)

▶

20

⏮

⏭

⚙

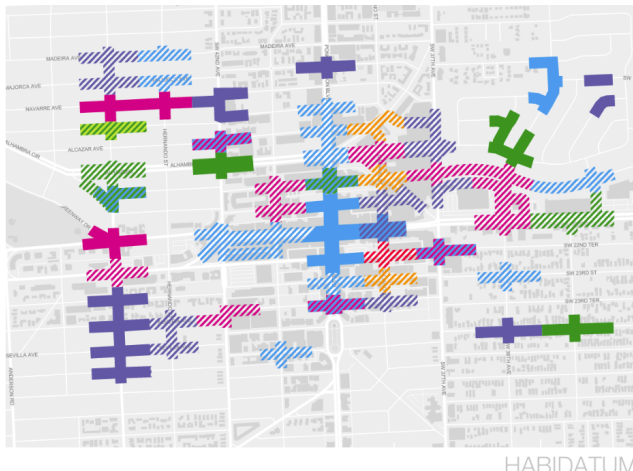


Рис.: Habidatum, урбанистика