BASIC SYNTACTIC RELATIONS FOR SENTIMENT ANALYSIS

R. Mavlyutov (arslan), N. Ostapuk (nataxane) @yandex-team.ru

Next: prerequisites

Prerequisites: reviews analysis

Review #33242

(+) Расположение превосходное.
(-) Все остальное - ужасно. В комнате было безумно холодно. Ремонт в этой комнате делали лет 20 назад, дверь в комнату открывалась с трудом. Конечно стоимость низкая, но мне доводилось останавливаться в более приличных условиях за меньшие деньги. Ванная - тихий ужас. Грязно и плитка отваливается.

Next: prerequisites

Prerequisites: parameters

Review #33242

(+) Расположение превосходное.
(-) Все остальное - ужасно. В комнате было безумно холодно. Ремонт в этой комнате делали лет 20 назад, дверь в комнату открывалась с трудом. Конечно стоимость низкая, но мне доводилось останавливаться в более приличных условиях за меньшие деньги. Ванная - тихий ужас. Грязно и плитка отваливается.

Next: prerequisites

Prerequisites: sentiments

Review #33242

(+) Расположение превосходное
(-) Все остальное - ужасно. В комнате было безумно холодно. Ремонт в этой комнате делали лет 20 назад. дверь в комнату открывалась с трудом. Конечно стоимость низкая, но мне доводилось останавливаться в более приличных условиях за меньшие деньги. Ванная - тихий ужас. Грязно и плитка отваливается.

Prerequisites: obj + sent

parameter	review	mark
расположение	превосходное	**+1**
комната	холодно	**-1**
ремонт	20 лет назад	**-1**
стоимость	низкая	**+1**
условия		**0**
ванная комната	ужас	**-1**

Next: prerequisites

Prerequisites: ways to express view

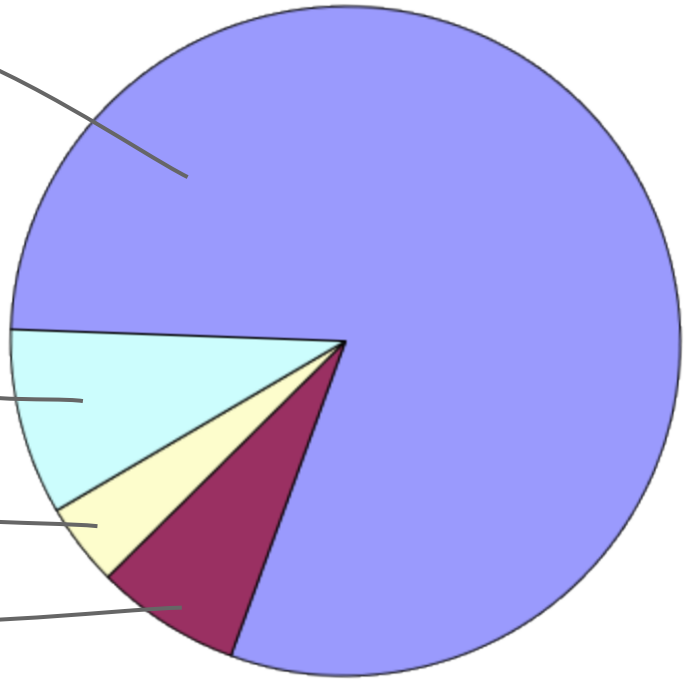
- train set: 2000 reviews
- subjective evaluation was expressed by:

adjectives - 80%

complex expressions - 9%

adverbials - 4%

predicates - 7%



Next: prerequisites

ROMIP 2013: Sentiment track, 2 classes

Next: prerequisites

Classifier: main ideas

- get a set of templates for sentiments
  positive: verb:понравиться -> adv:очень
  positive: adj:красивый -> noun:*

- find the templates in texts

- calculate weighted sum and make a decision

Next: prerequisites

Classifier: templates how?

Tomitaparser (api.yandex.ru/tomita/) for extraction of:

- object parameters
- basic syntactic relations:
  - preposition + noun group
  - adverb + adjective, verb, noun
    - очень красивый, работал хорошо
    - цены очень даже ничего
  - adjective + noun
    - хорошее обслуживание, обслуживание замечательное,
    - бармен (казался|был) немым
  - noun + noun
  - verb + noun - low quality, but useful!
- negations
  - не работает, без недостатков, ни одного человека, и т.п.

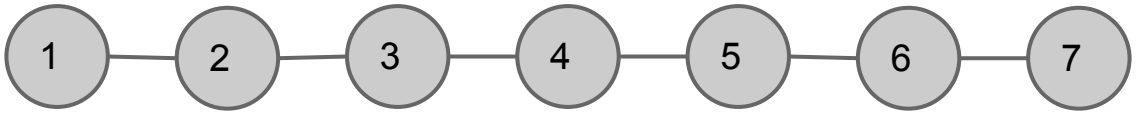
Next: prerequisites

Sentence graph: templates how?

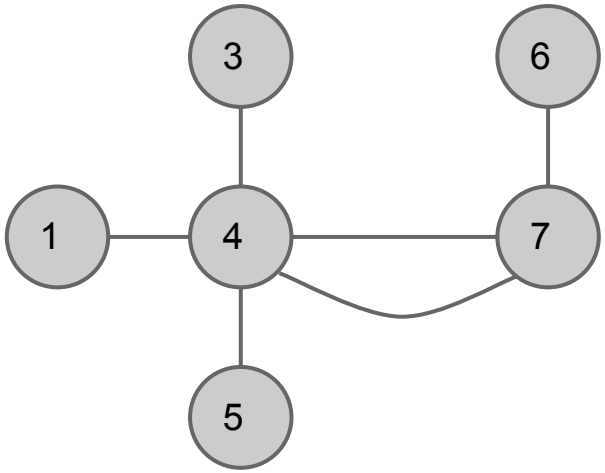
1 2 3 4 5 6 7

Спокойный и нетеропливый фильм Джармуша весьма порадовал

no syntax:



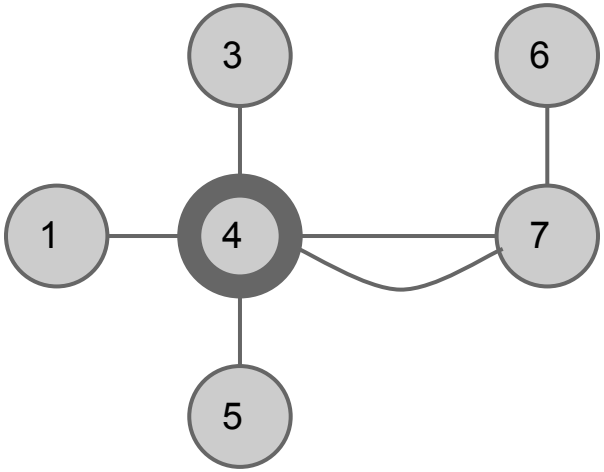
+ syntax:



фильм порадовал

Next: prerequisites

N-grams on graph



Bigrams:

1 - **4**

4 - 5

3 - **4**

4 - 7

Trigrams:

1 - **4** - 3

1 - **4** - 5

1 - **4** - 7

4 - 7 - 6

Next: prerequisites

N-grams unification

Triplet from the graph: Персонал - помог - очень

Персонал = [lemma:персонал, parameter, pos:noun, *]
Помог = [lemma:помогать, pos:verb, *]
Очень = [lemma:очень, pos:adv, *]

So, 1 triplet forms 36 n-grams:

lemma:персонал && lemma:помогать && lemma:очень
lemma:персонал && lemma:помогать && pos:adv
lemma:персонал && lemma:помогать && *
...
lemma:персонал && pos:verb && *
parameter && lemma:помогать && *
* && lemma:помогать && *
etc.

Next: prerequisites

Why syntax?

Ngram: впечталение&&[not_]произвести

Examples:

Фильм не произвел такого сильного впечатления...

Фильм произвел громадное впечатление.

Да, впечатление произвели.

Next: prerequisites

Classifier: main components

- templates (sentiments) dictionary

- object thesaurus

- technique for text transformation to connect sentiments and objects (sentence graph)

- some approach to make decision on the graph (ML?)

Next: prerequisites

Sentiments: sources

Sentiment = template, ngram (positive or negative opinion)

Common (not object):

- our old results (~1100)

  плохой, дурацкий, надоесть, грязь, трэш, гадко, not_то ("уже не тот"), noun&&ничто, noun&&оставлять_желать

- sentiments proposed by the organisation committee (+ ~400)
- expanded via linguistics rules (!not evaluated)

Specific (object-dependent or multiwords):

- mined: (LJ.com - Wikipedia.ru) x Train.set

  корпус люфтит, битый пиксель

Next: prerequisites

Sentiments: mining of complex

битый пиксель, синий экран смерти, ..

- PMI between the Training set and Wikipedia.ru

- SELECT * FROM train_set WHERE freq > 2 AND class_freq > 0.8

Sentiments: expansion by prefixes

Set: 200K articles from LJ.com

Algo:

- for each word with frequency > 10: find suffixes at least 5 characters length, what have a higher frequency. If it has, then use prefix.
- check each prefix, that was at least before 4 different roots

Result:

around 300 prefixes, filtered to 40:

- анти, без, во, воз, вос, все, вы, высоко, ..

не, недо, пере, псевдо, супер, мега. гипер

Next: prerequisites

Sentiments: expansion by roots

Set: 200K articles from LJ.com

Algo:

- lemmas with freq > 20: get all pairs: (root, suffix)
- join suffixes with same root
- for each suffixes pair count amount of roots they were together
- use pairs with freqs not less than 50

Result:

around 600 pairs like:

"" -> -ся, "" -> -ик, -о -> -ый,

-ность -> -ый, -ние -> -ть, -вание -> -вать, ...

Next: prerequisites

Sentiments: expansion results

Sentiment entries count increased on 79%
(in the 200K LJ.com)

Examples:

By prefixes:

счастливый - несчастливый, чистый - всечистый, терпеть - притерпеть, шедевр - микрошедевр, классный - супермегаклассный, умный - малоумный, (!bad) культурный - межкультурный

By suffixes:

суперпозитив -> суперпозитивный, безумство - безумствовать, влюблять - влюбляться, глючить -> глючок, глючка, тщеславие - тщеславность, вонючий - вонючка, погубительный -> погубить...

Next: prerequisites

Object thesaurus

Three types of objects:
- proper names (given)
 - common names (extracted by PMI*)
 - parameters (extracted by PMI*)

*Pointwise mutual information: $$\operatorname{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)}$$

2 stages:
1) books and films vs. cameras - via PMI
2) books vs. films - manually

Next: prerequisites

Object thesaurus: examples

- common names for cameras:

фотоаппарат, фотик, фотокамера, камера, мыльница, кодак, кэнон, лейка, сони, сонька, агрегат, аппарат, аппаратик...

- parameters for films:

костюм, концовка, мимика, актер, режиссер, изложение, финал, сцена, графика, детализация, задумка...

Next: prerequisites

Object thesaurus: results

	books	films	digital cameras
common nouns	69	74	475
proper names	2713	208	1412
parameters	161	252	512

In the training set: found something from the Object thesaurus in 84% of texts.

Next: prerequisites

Making decision

- only sentences close to objects from the thesaurus

- use more specific templates and templates with objects

- intuition: if everything is right, we don't need weights

- common and proper object mentions are more important than parameters

- linear model

- train using SVM on the train set

Next: prerequisites

Making decision: model

K - coefficient
N - count in text
Indexes: c+ - common name and positive sentiment
c- - common name and negative sentiment
p+ - parameter and positive sentiment
p- - parameter and negative sentiment
prop-, prop+
-, + - sentiment not linked to object

Prediction = SUM(Ki * Ni) > 0

Next: prerequisites

Results: books

Accuracy = 0.86, III place

Precision+ = 0.873 / 0.91
Recall+ = 0.982 / 0.935

Precision- = 0.33 / 0.58
Recall- = 0.058 / 0.411

Next: prerequisites

Results: films

Accuracy = 0.828, II place

Precision+ = 0.836 / 0.857
Recall+ = 0.979 / 0.949

Precision- = 0.682 / 0.604
Recall- = 0.192 / 0.333

Next: prerequisites

Results: mistakes

Шось я расписАлась в последнее время.... Дочитала " Азазель " Акунина. Не поняла откуда столько восторгов то, детектив обычный, есть писатели этого жанра гораздо интереснее. Люди, покупать продолжение или что то другое этого автора? Или интереснее не предвидится? Если что я из детективов люблю (ну не то шоп люблю, но могу читать с интересом) Агату Кристи, Рекса Стаута и Артура Хейли . А вообще: Макса Фрая, Лукьяненко, Хайнлайна, Бредбери... Насоветуйте мне книг, а? ЗЫ. Прочла 3 книги ЖЖ-шных авторов: " Манюня " Наринэ Абгарян " Сантехник, его кот, жена и другие подробности " Слава Сэ " Держите ножки крестиком, или Русские байки английского акушера " Денис Цепов Все понравились, над последней я иногда смеялась вслух. {HAPPY}

Thank you! :)

LiveJournal: a lot of extraneous info

Гарри Босх Майкла Коннелли. Гарри Босх - герой серии детективных романов Майкла Коннелли . Жанр напоминает нуар, только герой здесь полицейский и время другое. Другое, но тоже не слишком пригодное для проживания. Гарри - странный полицейский. Он не признает правил, неуживчив, одиночка, " последний койот " , как и называется один из романов. Система не может принять такого человека, каким бы результативным он ни был, она выпихивает его всеми способами, поэтому мой Гарри в каждом романе и победитель и обвиняемый. Хотя выпихнуть его окончательно не получается. Слишком он хитер, упрям и живуч. А иногда он так упрям, что даже раздражает. " Трудно тебе поступить по правилам?, - спрашиваю я его, - ты же опять огребешь " . " Ну и огребу, - отвечает он, - а ты думаешь, мне это так важно, как тебе? " Интересный цикл, сюжеты, детали, атмосфера. В начале романа время течет медленно, но потом все убыстряется, убыстряется и оторваться от книги ты уже не можешь.

!!! Links to objects are important

Next: prerequisites

LiveJournal: ways to express view

- set: 100 texts from test set (LJ.com)
- subjective evaluation was expressed by:

adjectives - 38%

predicates - 28%

nouns - 17%

complex expressions - 13%

они все могут, это нечто, новый уровень ужастика, притянутый за уши, ...

adverbials - 4%

Next: prerequisites

Training set: imhonet + market.yandex

- rich for sentiment unigrams

those that constitute 96% of all sentiments entries are in the training set with frequency > 5

- not large enough for connections (bigrams, trigrams, etc.)

- different types of text, no extraneous info

Next: prerequisites