Участники

- Алексей Виняр – главный по ручной разметке, обсуждение архитектуры системы, анализ результатов
- Даша Игнатенко – предобработка корпуса, работа со словарем RuSentiLex, анализ результатов
- Катя Герасименко – код, анализ результатов

1. Оценочные выражения

Перечислены в sentiment_words.txt.

76 слов, из них 17 – совсем ошибки (см. п. 2.4).

2. Отчет

Общая конструкция системы:

1) обучить w2v на наших текстах
2) найти кандидатов на оценочные слова с помощью w2v - найти most_similar, а потом most_similar от тех, что нашлись на первой итерации
3) с помощью NMF на наших текстах получить 20 тем
4) найти топ-10 тем, которые NMF присваивает словам из золотого стандарта
5) найти темы для кандидатов из шага 2 и оставить среди них те, у которых больше трех топовых тем - из топа тем из шага 4

2.1. Корпус

Мы взяли отзывы о ресторанах (SentiRuEval_rest_train.xml) и лемматизировали их с помощью Mystem (lemmas.txt). Дальше всё обучали на леммах. Объем: 3022287 токенов в 19034 документах.

2.2. word2vec

Модель на лемматизированных документах, 300 измерений, окно в 7 токенов, минимальная частота слова 3.

Как составляли список кандидатов на оценочные слова:

- лемматизировали слова из золотого стандарта, если это одно слово
- находили самые похожие слова – 10 по most_similar + ограничение на схожесть – не меньше 0.6 с исходным словом
- для каждого похожего слова – похожие на него – 10 по most_similar + ограничение на схожесть – не меньше 0.6 с исходным словом

Недостатки подхода:

1) не обрабатываются неоднословные оценочные выражения

2) не генерируются неоднословные оценочные выражения
3) порог схожести – гиперпараметр, который мы подобрали руками-глазами
4) много мусора (отдельный камень в сторону «вау» и «подошва» из золотого стандарта)

2.3. Topic modelling

- Много мусора из метода w2v => надо фильтровать
- Обучающего корпуса на расстоянии одного клика не оказалось => unsupervised learning
    - ⇨ Мы попробовали NMF и нам понравилось.

Процесс:

1) векторизация (Tf-idf) лемматизированных текстов – минимальная частота 10, выкинуть стоп-слова (список стоп-слов - из NLTK)
2) обучение NMF на полученной матрице – 20 тем
3) для всех слов золотого стандарта:
    a. векторизировать как документ
    b. найти 5 наиболее вероятных топиков для этого «документа»
4) выбрать 10 топиков, в которые чаще всего попадают слова из золотого стандарта
5) для каждого кандидата, полученного из word2vec:
    a. векторизировать как документ
    b. найти 6 наиболее вероятных топиков для этого «документа»
    c. если из них больше 3 топиков есть в топе-10 из п.4, то добавить его в конечный список

Недостатки:

1) большое количество гиперпараметров – минимальная частота слова для векторизатора (10), количество тем (20), топ для оценочных слов (10), топ для каждого документа (5 для ЗС, 6 для кандидатов – для менее жесткой фильтрации). Подбирались руками-глазами
2) и все равно выкинулись хорошие слова. Но при уменьшении жесткости фильтрации в конечный список попадало много мусора

2.4. Оценка

Оценка – доля слов из нашего списка, которые есть в словаре RuSentiLex-2017 (http://www.labinform.ru/pub/rusentilex/index.htm).

Из наших 76 слов только 33 есть в этом словаре (0.434), 43 остались за бортом. Но надо посмотреть на них глазами.

действительно не оценочные	**содержат какую-то долю оценки**
1) out	1) благодарить
2) андрей	2) великолепно
3) барбареско	3) вкусно
4) гардеробщик	4) жеваться
5) громкость	5) завуалировать

6) девушка
7) жир
8) картошка
9) маринованный
10) обслуживающий
11) общаться
12) ольга
13) разговаривать
14) раздевать
15) слушать
16) смородиновый
17) собеседник

6) зажигательный
7) зажигать
8) замечательно
9) негромкий
10) ненавязчивый
11) обалденный
12) отблагодарить
13) отвратительно
14) перекрикивать
15) поздороваться
16) порадовать
17) приветливо
18) разочаровывать
19) ребята
20) удручающий
21) улыбка
22) умничек
23) услужливый
24) учтивый
25) шутить
26) этнический

Несколько комментариев:

1) *громкость* будет скорее всего там, где она была либо приятнее ожидаемой, либо неприятнее (выше), то есть упоминаться там, где есть оценка.
2) *жир*, конечно, само по себе не оценочное, но в отзыве скорее будет употребляться для негативного описания еды.
3) *собеседник* тоже может быть связан с оценкой громкости, поэтому попал в автоматический список.
4) *ребята* в отзывах содержат в себе оценку – либо положительную, либо уменьшительно-ироническую.
5) *умничек* – это неправильная лемматизация слова «умнички».
6) *этнический* считаем словом, содержащим в себе оценку, скорее положительную.
7) *поздороваться* и *шутить* используются для характеристики персонала, поэтому их можно рассматривать как содержащие оценку, но это довольно спорный случай.
8) В словаре многих из этих слов нет, потому что они не являются оценочными сами по себе, но в данном контексте содержат оценочность.