**Проект**

**Выполнили: Зеленкова Валерия, Суслова Дарья, Иванов Вячеслав.**

Лемматизированный корпус и модель W2V находятся в папке: https://drive.google.com/open?id=1yeIL9BOr3pWiueLKR6tFipDuUImmLfNU

1.Определение оценочной лексики с использованием seed (золотого стандарта) и W2V.

Для для этой части нашего проекта был взят Word2Vec и список ЗС. Из списка были удалены коллокации и оставлены ярко оценочные слова (слово «подошва» было убрано):

seed = ['вежливый\_A', 'великолепный\_A', 'вкусный\_A',

'внимательный\_A', 'большой\_A', 'выбор\_S', 'вау\_INTJ', 'неплохой\_A',

'идеальный\_A', 'неприятный\_A', 'понравиться\_V', 'рекомендовать\_V',

'советовать\_V', 'отвратительный\_A', 'отличный\_A', 'положительный\_A',

'отрицательный\_A', 'приятный\_A', 'радовать\_V', 'молодец\_S',

'улыбаться\_V', 'приветливый\_A', 'отличный\_\_A', 'замечательный\_A']

Затем в модели проверили наличие слова, нашли для каждого из них 10 ближайших соседей и записали их в массив, чтобы затем найти их в отзывах.

Перед поиском в отзывах у слов были удалены маркеры частей речи, которые приписываются моделью. Далее была проведена предобработка текстов (лемматизация) для того, чтобы проверить наличие найденных слов в текстах отзывов.

Был получен расширенный список. Однако, минусы этого списка в том, что некоторые слова, попавшие в него, не являются оценочными. Это, например, синонимы слова "молодец": парень, хлопец. Также, как нам кажется, случайно попали слова: дивчина, ухмыляться, подмигивать, заулыбаться, засмеяться, рассмеяться, посмеиваться, хмыкнуть, нужно, просить, упрашивать, вынуждать, приходиться, намерен. Поэтому стоит их удалить. Так выглядит расширенный список из 177 слов в итоге: 'доброжелательный', 'сдержанный', 'приветливый', 'тактичный', 'любезный', 'учтивый', 'обходительный', 'ласковый', 'корректный', 'деликатный', 'превосходный', 'прекрасный', 'изумительный', 'роскошный', 'замечательный', 'отличный', 'дивный', 'отменный', 'бесподобный', 'чудный', 'аппетитный', 'сытный', 'калорийный', 'невкусный', 'сладкий', 'вкусно', 'ароматный', 'превкусный', 'наваристый', 'вкуснющий', 'вдумчивый', 'доброжелательный', 'недоверчивый', 'пристальный', 'проницательный', 'зоркий', 'невнимательный', 'ласковый', 'благожелательный', 'настороженный', 'огромный', 'громадный', 'небольшой', 'немалый', 'колоссальный', 'значительный', 'изрядный', 'больший', 'малый', 'незначительный', 'выбирать', 'усмотрение', 'самовыражение', 'предоставлять', 'передвижение', 'одолжение', 'предоставляться', 'позиционирование', 'отличный', 'хороший', 'недурной', 'приличный', 'превосходный', 'замечательный', 'отменный', 'плохой', 'посредственный', 'солидный', 'иллюзорный', 'целостный', 'абстрактный', 'реальный', 'оптимальный', 'приятный', 'тягостный', 'странный', 'малоприятный', 'нехороший', 'тревожный', 'щекотливый', 'обидный', 'волнующий', 'забавный', 'нравиться', 'обрадоваться', 'полюбиться', 'заинтересовывать', 'обрадовать', 'порадовать', 'удивлять', 'огорчать', 'довольный', 'советовать', 'порекомендовать', 'рекомендоваться', 'посоветовать', 'предлагать', 'настаивать', 'предлагаться', 'предписывать', 'поручать', 'утверждать', 'посоветовать', 'рекомендовать', 'рекомендоваться', 'предлагать', 'мерзкий', 'омерзительный', 'ужасный', 'безобразный', 'гнусный', 'скверный', 'постыдный', 'жуткий', 'пошлый', 'отвратный', 'отменный', 'превосходный', 'неплохой', 'хороший', 'великолепный', 'замечательный', 'плохой', 'первоклассный', 'недурной', 'улучшаться', 'отрицательный', 'позитивный', 'негативный', 'объективный', 'положительный', 'негативный', 'позитивный', 'неприятный', 'привлекательный', 'отрадный', 'интересный', 'приятно', 'грустный', 'волнующий', 'симпатичный', 'радостный', 'занятный', 'удивлять', 'восхищать', 'смущать', 'волновать', 'умилять', 'порадовать', 'обрадовать', 'раздражать', 'веселить', 'огорчать', 'молодчина', 'умница', 'дюжий', 'здорово', 'спасибо', 'ласковый', 'доброжелательный', 'добродушный', 'улыбчивый', 'вежливый', 'обходительный', 'обаятельный', 'общительный', 'застенчивый', 'выдающийся', 'гениальный', 'знаменитый', 'превосходный', 'талантливый', 'изумительный', 'великолепный', 'удивительный', 'хороший', 'прекрасный'

Затем был построен список коллокаций со словами из финального списка: [('очень', 'вкусно'), ('остаться', 'довольный'), ('отдельный', 'спасибо'), ('приятно', 'удивить'), ('огромный', 'спасибо'), ('очень', 'приятный'), ('очень', 'довольный'), ('спасибо', 'за'), ('очень', 'нравиться'), ('очень', 'хороший'), ('положительный', 'эмоция'), ('отличный', 'место'), ('спасибо', 'вы'), ('очень', 'приятно'), ('не', 'плохой'), ('огромный', 'благодарность'), ('приятный', 'впечатление'), ('приятный', 'атмосфера'), ('желать', 'хороший'), ('большой', 'спасибо'), ('очень', 'порадовать'), ('вкусно', 'и'), ('самый', 'хороший'), ('хороший', 'место'), ('вкусно', 'поесть'), ('вкусно', 'но'), ('я', 'нравиться'), ('с', 'больший'), ('положительный', 'отзыв'), ('вполне', 'приличный'), ('весь', 'советовать'), ('огромный', 'количество'), ('больший', 'количество'), ('вежливый', 'и'), ('очень', 'приветливый'), ('неприятный', 'осадка'), ('приветливый', 'персонал'), ('выразить', 'огромный'), ('не', 'советовать'), ('настоятельно', 'рекомендовать'), ('очень', 'вежливый'), ('хороший', 'чем'), ('небольшой', 'компания'), ('неплохой', 'но'), ('быть', 'приятно'), ('приятно', 'провести'), ('приятный', 'музыка'), ('приятный', 'место'), ('самый', 'интересный'), ('приятный', 'интерьер'), ('положительный', 'впечатление'), ('долго', 'выбирать'), ('негативный', 'отзыв'), ('приятный', 'обстановка'), ('хороший', 'настроение'), ('замечательный', 'место'), ('спасибо', 'администратор'), ('небольшой', 'количество'), ('хороший', 'обслуживание'), ('очень', 'интересный'), ('спасибо', 'огромный'), ('только', 'положительный'), ('прекрасный', 'место'), ('приятный', 'аппетит'), ('самый', 'приятный'), ('интерьер', 'приятный'), ('за', 'прекрасный'), ('приятно', 'удивлённый'), ('отличный', 'кухня'), ('это', 'замечательный'), ('приветливый', 'и'), ('советовать', 'посетить'), ('очень', 'здорово'), ('хороший', 'смысл'), ('ничто', 'выдающийся'), ('особенно', 'порадовать'), ('весь', 'рекомендовать'), ('просто', 'замечательный'), ('спасибо', 'весь'), ('ничто', 'плохой'), ('отличный', 'обслуживание'), ('персонал', 'приветливый'), ('спасибо', 'большой'), ('вкусно', 'покушать'), ('огромный', 'плюс'), ('приятный', 'сюрприз'), ('спасибо', 'повар'), ('замечательный', 'ресторан'), ('приличный', 'место'), ('вежливый', 'персонал'), ('какой-то', 'странный'), ('хороший', 'ресторан'), ('ожидать', 'больший'), ('намного', 'хороший'), ('не', 'раздражать'), ('огромный', 'удовольствие'), ('самый', 'плохой'), ('из', 'хороший'), ('хороший', 'но'), ('мера', 'сладкий')]

А также мы посмотрели, сколько пересечений со словарем Sentilex из нашего расширенного списка: совпадений 101 слово.

**Код с комментариями - в ipynb файле Project\_with\_SEED**

2. Определение оценочной лексики с помощью Non-Negative Matrix Factorization (извлечение ключевых слов)

С помощью NMF был извлечен ряд ключевых слов. Наряду с этим вручную было записано 19 независимых оценочных базовых слов, векторы которых были затем найдены в pre-trained модели word2vec. Затем каждое слово из ключевых мы сравнили с каждым словом из независимого списка, установив порог дальности cosine distance < 0.9.

Так получился следующий список (всего 93 слова):

*'адекватный\_A наверное\_ADV новое\_S низкий\_A единственный\_A соответствовать\_V удобный\_A прекрасный\_A радовать\_V приемлимый\_A рекомендовать\_V интересный\_A всякий\_APRO порадовать\_V добрый\_A большой\_A приемлемый\_A оригинальный\_A вкусно\_ADV положительный\_A приличный\_A написать\_V гуманный\_A плохо\_ADV обходительный\_A общий\_A уважаемый\_A великолепный\_A отвратительный\_A стильный\_A классный\_A пожалеть\_V красивый\_A нормальный\_A отзывчивый\_A молодой\_A испортить\_V идеальный\_A комфортно\_ADV вообще\_ADV довольный\_A выбирать\_V главное\_S расторопный\_A неплохой\_A хороший\_A разнообразный\_A завышать\_V приятно\_ADV удивлять\_V обычный\_A достойный\_A ужасный\_A доброжелательный\_A замечательный\_A настроение\_S частенько\_ADV приятный\_A дорого\_ADV качественный\_A испорченный\_A нравиться\_V хотеть\_V хромать\_V шикарный\_A отличный\_A неплохо\_ADV отменный\_A оказываться\_V демократичный\_A желать\_V рождение\_S необычный\_A хотеться\_V вкус\_S вкусный\_A качество\_S фоновый\_A умеренный\_A мило\_ADV плохой\_A понятно\_ADV дружелюбный\_A благоприятный\_A навязчивый\_A понравиться\_V здорово\_ADV высокий\_A вежливый\_A работать\_V любимый\_S видимо\_ADV чудесный\_A'*

Видно, что здесь действительно очень много оценочной лексики, но есть и не очень подходящие слова (например, “вкус”). Мы посмотрели, сколько пересечений со словарем Sentilex из этого списка:

found\_in\_sentilex = 49

Кроме того, был построен список коллокаций со словами из финального списка:

*[('в', 'общий'), ('день', 'рождение'), ('очень', 'понравиться'), ('молодой', 'человек'), ('очень', 'вкусно'), ('очень', 'вкусный'), ('оставаться', 'довольный'), ('высокий', 'уровень'), ('приятно', 'удивлять'), ('хотеться', 'бы'), ('очень', 'приятный'), ('мой', 'любимый'), ('всякий', 'похвала'), ('не', 'понравиться'), ('очень', 'довольный'), ('понравиться', '.'), ('не', 'пожалеть'), ('очень', 'хороший'), ('высоко', 'всякий'), ('не', 'хотеться'), ('самый', 'главное'), ('вкусно', ',\n'), ('общий', ',\n'), ('очень', 'нравиться'), ('вкусный', 'еда'), ('по', 'вкус'), ('на', 'вкус'), ('большой', 'количество'), ('я', 'понравиться'), ('оставлять', 'желать'), ('общий', 'впечатление'), ('добрый', 'день'), ('на', 'высокий'), ('вкусно', '.'), ('единственный', 'минус'), ('положительный', 'эмоция'), ('наверное', ',\n'), ('отличный', 'место'), ('.', 'единственный'), ('вкусный', ',\n'), ('очень', 'приятно'), ('большой', 'спасибо'), ('большой', 'компания'), ('не', 'плохой'), ('фоновый', 'музыка'), ('желать', 'хороший'), ('всякий', 'случай'), ('понравиться', ',\n'), ('приятный', 'впечатление'), ('самый', 'вкусный'), ('со', 'вкус'), (',\n', 'наверное'), ('мой', 'вкус'), ('вкусно', '!'), ('приятный', 'атмосфера'), (',\n', 'видимо'), ('самый', 'хороший'), ('очень', 'порадовать'), ('хотеть', 'выражать'), ('в', 'качество'), ('хотеть', 'отмечать'), ('видимо', ',\n'), ('хороший', 'место'), ('порция', 'большой'), ('все', 'понравиться'), ('вкусно', 'поесть'), ('довольный', '.'), ('не', 'плохо'), ('не', 'хотеть'), ('я', 'нравиться'), ('не', 'навязчивый'), ('вообще', 'не'), ('хотеть', 'сказать'), ('вполне', 'приличный'), ('вкусно', 'и'), ('положительный', 'отзыв'), ('очень', 'красивый'), ('очень', 'достойный'), ('большой', 'выбор'), ('любой', 'вкус'), ('хотеться', 'отмечать'), ('качество', 'еда'), ('не', 'соответствовать'), ('написать', 'отзыв'), ('еда', 'вкусный'), ('испортить', 'настроение'), ('вкусный', '.'), ('вкус', '.'), ('большой', 'порция'), ('хотеть', 'поблагодарить'), ('вежливый', 'и'), ('удобный', 'диван'), ('а', 'главное'), ('хороший', ',\n'), ('неплохой', ',\n'), ('самый', 'интересный'), ('понятно', ',\n'), ('замечательный', 'место'), ('вкусовой', 'качество'), ('настоятельно', 'рекомендовать')]*

**Код с комментариями - в ipynb файле Project\_with\_topic\_modelling**

**Сравнение методов:** доля по отношению к кол-ву найденных слов в первом методе - 57% (совпало со словарем 101 слово из 177), а во втором - 52% (совпало со словарем 49 слов из 93)