**У проекта нет сформулированной темы.**

**Моя идея - выявить тематическое поле «степь» (т.е. с какими словами сочетается) для дальнейшего анализа материала (изучения концепта «степь») в рамках научной деятельности на кафедре.**

**Тексты лежат здесь:** <https://drive.google.com/drive/folders/1mYm-PKjNU6tK5sgbS5J2kT-phFwLKGOJ?usp=sharing>

**Файл Steppe\_Freq&Collocations**

В первой части я определяю частоту вхождений слова «степь» (без производных) в каждый текст и частоту слова «степь» относительно длины текста.

Корпус текстов разбит на две части: тексты оренбургский авторов или об уральских степях (всего 53 текста общим объемом 1 065 985 слов) и тексты иных авторов (всего 5 текстов общим объемом 810 801 слово).

В результате получены таблицы частот, на которых будет сделана визуализация в R.

Во второй части осуществлен поиск коллокаций со словом «степь» и его производными. Для этой цели в обоих подкорпусах текстов с помощью регулярных выражений отобраны предложения, содержащие искомые слова, причем важно было, чтобы в выборку не попали слова типа «Степан» и «степень». *(В регулярном выражении в настоящий момент есть выбор заглавной или строчной буквы, хотя все слова уже приведены к нижнему регистру, специально не изменяю регулярку, чтобы можно было применять ее на необработанных текстах)*.

Далее осуществлен поиск биграмм с использованием метрик raw\_freq, likelihood\_ratio и pmi. По такому же принципу можно найти триграммы, сравнивать и обобщать. Для подсчета частотности тех или иных биграм был применен морфологический парсинг, слова были приведены к нормальной форме, поэтому может показаться, что выражения типа «степь хутор» бессмысленны, но на самом деле это «степной хутор». Получены также таблицы частот для биграмм (файл steppe\_results).

**Файл Steppe\_Syntax**

В данном файле применен еще один способ найти коллокации с искомым словом с использованием синтаксического парсинга (ConlLu и UdPipe). Этот способ мне нравится больше, т.к. при поиске биграмм иногда получаются бессмысленные сочетания, типа «ваш степной» или «степь далеко». С использованием семантического парсера были выявлены словосочетания типа «adj + noun (степь)», «adj (степной) + noun», «object (степь) + verb». Других типов словосочетаний быть не может, т.к. в словообразовательном гнезде слова «степь» нет глаголов.

На основе полученных данных уже можно делать вывод о концептуальном восприятии слова «степь» авторами художественных текстов.

В промежуточной проекте мной был использован другой способ поиска коллокаций с помощью морфологического парсинга, он тоже давал достаточно хороший результат, однако в данных было много мусора, поэтому захотелось разобраться с UdPipe.

**Файл Steppe\_clustering**

С помощью модели LSA выполнено тематическое моделирование предложений их художественной и из научной литературы со словом «степь». Всего в подкорпусе предложений из художественной литературы 1394 предложений, в подкорпусе предложений из научной литературы 1234 предложения.

Поскольку использовались только предложения с искомым словом, то результаты тематического моделирования на мой взгляд оказались смазанными, т.е. в них нет каких-то особых четко обозначенных тем. Немного лучше ситуация с научными текстами, там можно относительно говорить о темах.

(В планах на будущее взять не предложение, а абзац с искомым словом. Возможно, разбиение по темам будет более четким.)

**Файл Steppe\_classification**

По тетрадкам, которые мы изучали на ML, была выполнена классификация по наивному Байесу и SVM. Общий объем датасета – 2628 единиц.

Тетрадь не доработана, т.к. с некоторыми вопросами я пока не разобралась, например, как дать модели наивного байеса совершенно новый датасет.

**Из того, что можно сделать еще:**

- сравнить коллокации с тем, что дает НКРЯ

- использовать векторную семантику и найти ближайших соседей слова «степь»

- брать абзацы со словом «степь» и вытаскивать оттуда все существительные, считать частоту и строить тематическое поле с ядром и периферией