### Папка **code** содержит код, использованный для извлечения текстов, очистки и получения эмбеддингов

1. **tyv\_wiki\_parsing** - парсинг википедии на тувинском языке
2. **енм\_preprocess** - очистка данных (здесь указаны ссылки на статьи с источники, использованных при предобработке: стемминг, стоп-слова)
3. **tyv\_embeddings** - получение словаря эмбедингов
4. **Text2Vec** - получение текстов в эмбеддингах

### Папка **data** сожержит данные, полученные из википедии

1. **cat\_tyv\_links\_wiki** - содержит ссылки на все категории тувинской википедии
2. **all\_links** - содержит все ссылки на финальные страницы тувинской википедии, то есть на сами статьи с текстом
3. **clean\_texts** - очищенный литературный корпус вида <название, текст\n>. Очистка на данном этапе заключается в удалении пунктуации и инородных языков.
4. **clean\_texts\_final** - очищенный литературный корпус вида <название, текст\n>. Очистка на данном этапе заключается в стемминге слов и удалении стоп-слов
5. **tyv\_wiki\_full\_dict** - словарь эмбеддингов вида <слово: эмбеддинг>
6. **tyv\_wiki\_embedded\_texts** - список текстов, переведенных в эмбеддинги