Exploratory data analysis

Для качественного исследования и хорошего результата требуется знание о данных.

Мы имеем следующие наборы данных:

1. overall_80K.csv - датасет из параллельных предложений рус↔манси, предоставленный кейсодателем.
2. mansi.csv - собранный нами монокорпус из Луима Сэрипос.
3. dict.csv - распарсенный нами мансийско-русский словарь.
4. Синтетические данные.

Начнем по порядку.

overall_80K.csv

Данные параллельного корпуса, предоставляемые кейсодателем. Имеет 81,146 пар предложений рус↔манси.

A. От кейсодателя была получена информация о наличии кейсов, где слова написаны через пробел (пример: Г Е Р М А Н И Я). Для начала мы решили проверить количество этих кейсов используя регулярные выражения. Итого у нас вышло 847 различных предложений с данным дефектом. Данные кейсы были удалены из-за невозможности восстановить автоматически предложение.

   Из примеров видно, что далеко не всегда можно определить сколько слов написаны в таком стиле, чтобы правильно преобразовать предложение.

   Примеры:

Тӧ н т й и с м ӧ й т ы т Ю в а н Н и к о л а е в и ч такви щёпитасанэ.
Пӑвлув коныпал ты ляпат ӧлнэ мӑнь пӑвлытныл Нумто, Юильск ос сӑлыӈ мӑхум нӈвраманыл тыг Касумн т от ы гл ы я н ы л .
Т у в ы л ӑ г м ы ӈ м о с ы ӈ нӈврамыт ёт рӯпитэв.
Ань Урай ӯсныл ос К о н д и н с к и й р а й о н ы л ӧ л н э м ӑ х у м в ӑ т и х а л м ӑ н ь щ и пӑвылт ӯщлахтэгыт.
Маснут ӈнтуӈкве с а к а а т ӑ л ы м ё г ы т.
Мӧлты тӑл «Сургутнефтегаз» 1 5 х а н т ы н ӈ в р а м ӯщлахтын мӑгсыл олн тӑстыглас.
А к в а н а т ы м ол н ы т ул вос тӯлмантавет, м о р и ул в о с х ол т а в е т, о б щ е с т в е н ный организацият ты рӯпата вӑрнэ мӑхум ӧс уральтаӈкве патыяныл.
Т у в ы л И в д е л ь ӯ с мус машинал минм ы г т а с м ё н .
Ты т ӑ л с а к а с ӑ в хӧтпа яныт лавес, тӑн халанылт сӑв мӑньлат хӧтпа ӧлыс.

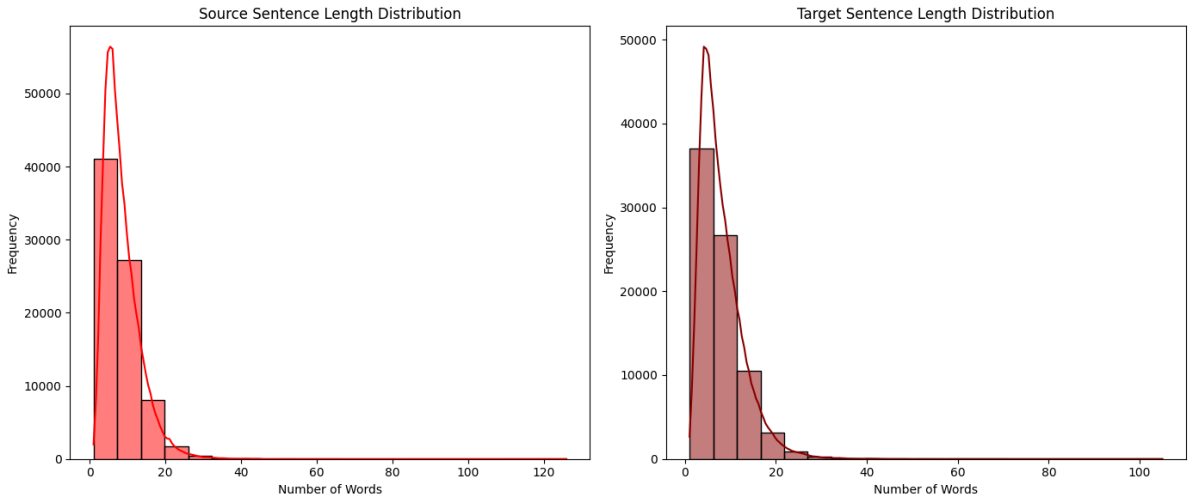
B. Значений NaN в датасете нет, есть полные дубликаты - их 251. Дубликаты были удалены, оставив только 1 версию.

C. Средняя длина предложений на мансийском составляет 8 слов, а максимальная целых 188, что слишком сильно разнится со средней длиной. Оказалось, что текст просто не был разбит на предложения:

“34. Ирина Константиновна Поята Хāльӯс район Лӧпмус пāвылт самын патыс. 35. Āще Константин Корнилович, оматэ Ольга Максимовна Албиныг ӧлсы̄г. 36. Экваг-ӧйкаг колтāглэнт китхуйплов няврам янмалтасы̄г. 37. Ань сāт хӧтпа хультыс.38. Ирина школа āстламе юи-пāлт Салехард ӯс медучилищан ханищтахтуӈкве минас. 39. Āстламе юи-пāлт Ямал мāн Тарко-Сале ӯс пӯльницан рӯпитаӈкве кӗтвес. 40. Тот мощ рӯпитас ос хум вāрыс. 41. Ӧйкатэн Молдавия мāн ӧлуӈкве тотвес. 42. Тувыл 1990 тāлт тэн ювле Хāльӯсн щӗмьяӈ тāгыл вāнтлысы̄г. 43. Нэ пӯльницат терапевт-лӗккарн нӗтым нёловхуйплов тāл рӯпитас. 44. Та юи-пāлт ос физиотерапия вāрмаль щирыл ӛлаль ханищтахтас. 45. Ань ты пӯсмалтан вāрмаль щирыл Хāльӯс пӯльницат китхуйплов тāл рӯпиты.46. Ирина Константиновна ӧйкатэнтыл кит няврам янмалтасы̄г. 47. Пыгӗн юридический академия āстлас. 48. Ань Хāльӯст ӧлы. 49. Āгитэн ос Ханты-Мансийск ӯс педколледжит нилыт тāл ханищтахты. 50. Мāн маньщи щӗмьят йильпи тāл кастыл янытлыянув!51. Ӧлупсанын кӯпнитыг вос ӧлы. 52. Йильпи тāлт нāн щуниӈыг вос ӗмтэгын.53. Рӯтанын, юртанын ёт сāв тāл пустāгыл ӧлэн.”

И это не единичный случай. Всего таких кейсов около 463х. Для всех этих кейсов с помощью регулярных выражений были убраны порядковые номера, а для предложений, чья длина больше 128, были разбиты на части поменьше. Кол-во предложений, чья длина больше 128 - 1.

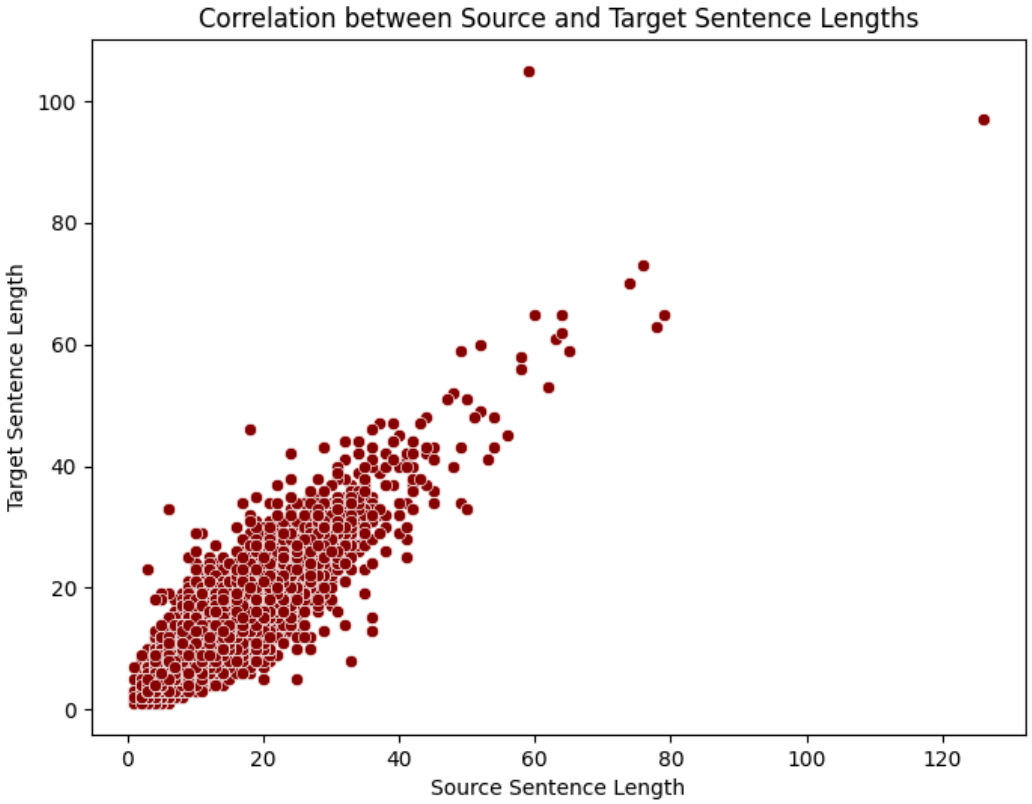
D. Теперь посмотрим на распределение длин предложений.



Из bar chart’a видно, что распределение похоже на экспоненциальное или

Хи-квадрат. Больше всего, конечно же, коротких предложений.

E. При визуализации корреляции длин источника и таргета было замечено сильное расхождение между семплами:

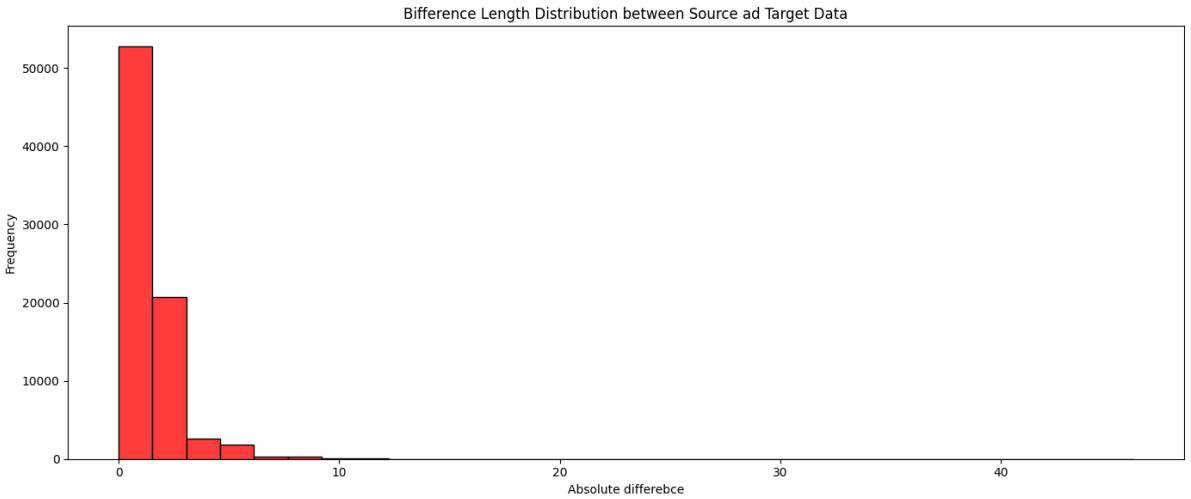



```
]:  differ.describe()
```

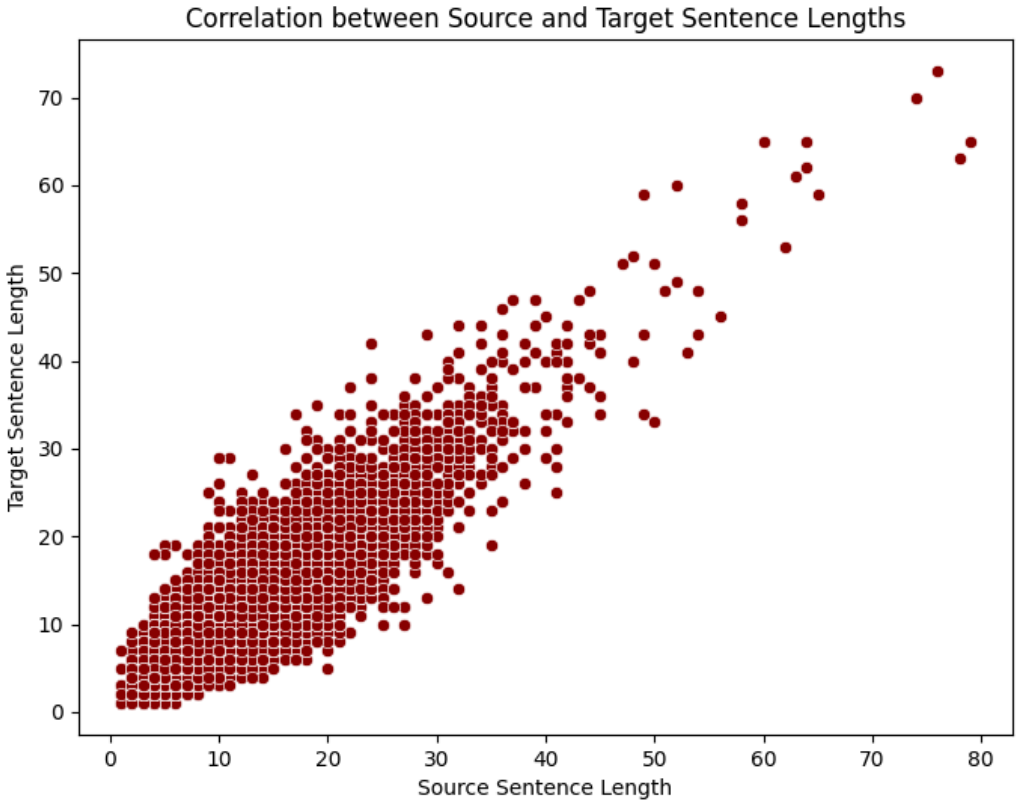
```
count    78702.000000
mean         1.322914
std          1.447477
min          0.000000
25%          0.000000
50%          1.000000
75%          2.000000
max         46.000000
dtype: float64
```

Был установлен трешхолд на разницу в длину равной 25. Видим следующие примеры:

Target	Source
Ӯлпыл , хӧлтан ӑлпылнува, ӯнтимен хӑпын - ам туп тарм, Маруся пӧсум сӧртнутын, Светлана ӑквьёт мӣннэ хӧтпаг - и нӑтэв ӑл тув, хӧтталь ӛри, лӑвегыт, ӑныг вӧр, хот ӑнымэгыт ӑ вӑтат кӣтыг ӯми хӑльиг, хӧт хӧнтыс мӧл- хӧтал кӧл ӧнтсыл ӧлнэ ӑгирись хӯрум ӗмас ӑӈк лӑхс.	Отпустил коня Иванушка - дурачок и взял с него слово - пшеницы больше не есть и не топтать.
Тувле кос ёмантас, луве лӑви: "Ул минэн, тыг йиен! Наӈ тах хотум мори ты номылматэгн, ӛква ты сайкалы, ӧлум пасмен тый; ам нёлссам асагумн сялтэн, тот капак ӧлы, та капакта хурум сярка аен!"	Конь говорит: "Не уходи, иди сюда!"
купса Сибиряков (Щипрах) лёнх оньщас - "Щипрах лёнхыг" (Сибиряковскии тракт) лавыглавес. Ялпын хоталт ёхталас Халь кс кущай Фомин В.И., Хулюмсунт миркол кущай Ануфриев Я.В. Янитлавест махум хотит олсыт н1лсат арыгкем тал. иильпи самын патум няврамыт. Янитлавест акван хасхатам (олмыгтам) махум. Сав потыртавест хоти махум олэгыт хоса аквёт (пуриянып- иив, келп, аргин, сорни). Няксимволь урыл эрыг хансум эква янитлавес. Няксимволь миркол кущай Волклва Т.К., потыртас, латын лавыс-хоти махум нетсыт ялпын хотал варункв. Няксимволь сав сунсылтаве Михаил Заплатин, Лев Вахитов - кинат, Игошев картинат мот хон мат тотыглавет, музей Ханты-Мансииск олэгыт. карыс сип ватат- павлув,тагт я овтохти (хайти витэ). Олнэ вармалюв минанти, та олантев - павылрисювт!	Первым поселением была манси деревня. потом из-за урала в поисках лучших мест для жилья пришли коми-зыряне. Из воспоминаний Е.М.Носовой, которая была в числе первых переселенцев она рассказывала: шли из-за Урала от голода, надеялись на рыбные места, пушнину. Мы выжили благодаря Няксимволю. Сосьвинская пристань Сибирякова - это единственное в данной местности поселение русский пункт, отстоящии от Березово в 500 верстах.



Было замечено, что предложения не были переведены до конца, а только первые части. Было сделано так: если target>source, то тогда берем первое предложение, иначе оставляем как есть. После данных манипуляций можем наблюдать следующую картину:



F. Также был проведен анализ частотности слов. Для русских предложений была проведена очистка от стоп слов и получилась такая картина:

Для мансийского языка мы не знаем стоп-слова, поэтому решили отобразить напрямую частотность слов:

Видно, что здесь много “коротких” слов. Мы связались со специалистами. знающие мансийский язык и спросили их о наличии предлогов. Ответ говорил о том, что предлогов в мансийском нет, но есть превербы и послеслоги, скорее всего это они и есть в бОльшем случае.

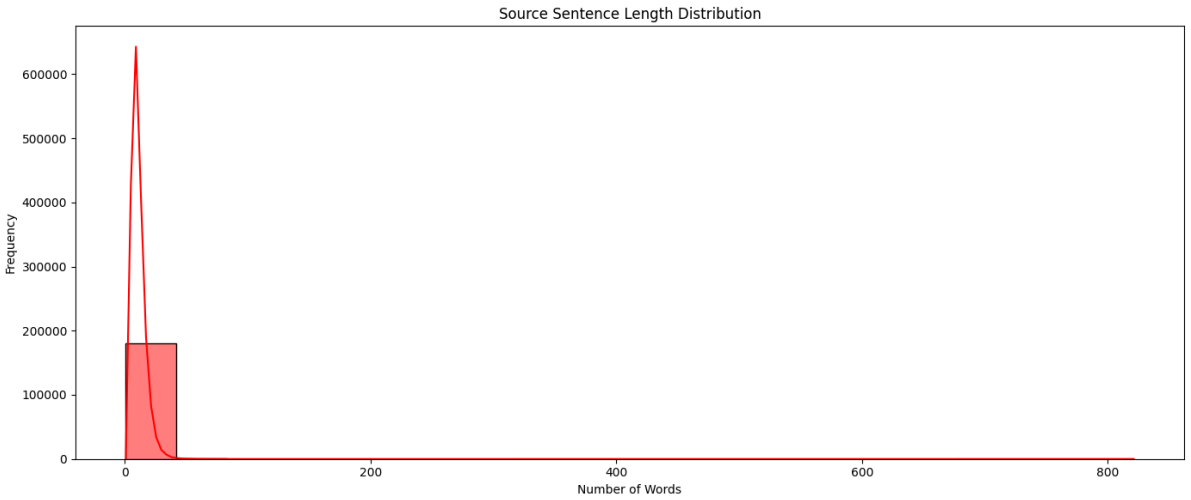
mansi.csv

Данный документ представляет из себя собранный монокорпус новостей на мансийском с сайта [Луима Сэрипос](), а также данные ученого [Csilla Horvath](). Использовался для обучения модели m2m100-418 на задаче mlm и в последующем

для генерации синтетических данных для дообучения модели.

A. Датасет разбит на 183124 предложения с помощью модуля nltk, предобработан тем, что убран спец символ ‘\xad’. Таблица имеет 2670 дубликатов, которые были убраны.

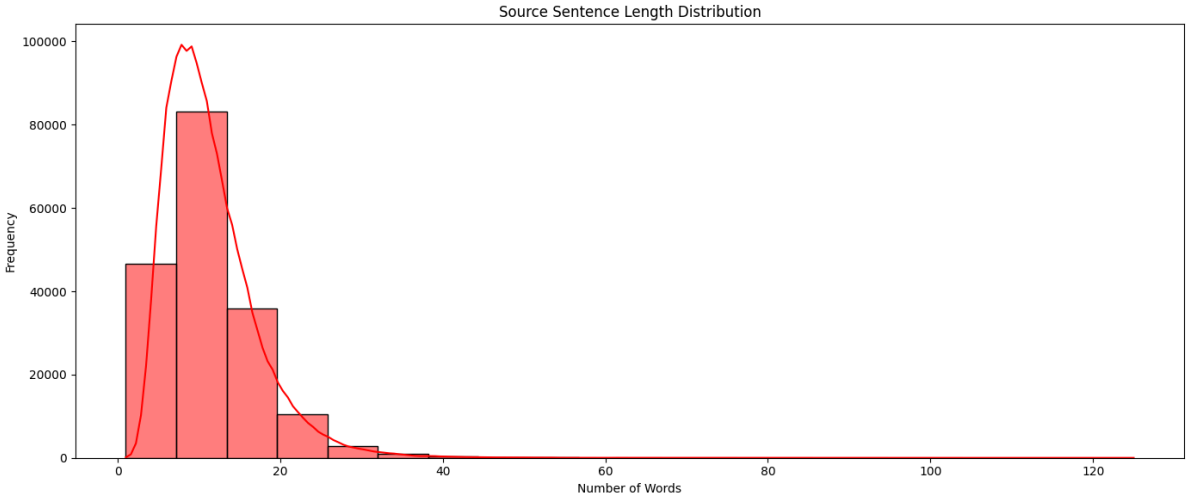
B. Датасет имеет следующее распределение длин предложений:




```
Summary of Mansi Text Lengths:
count    183124.000000
mean         11.417864
std           6.569508
min           1.000000
25%           7.000000
50%          10.000000
75%          14.000000
max         821.000000
Name: mns_length, dtype: float64
```

Предложение длиной в 821 слово действительно является одним предложением, так решили авторы статьи. Кол-во предложений с длиной больше 128 - 13.

Итоговое распределение для предложений, чья длина меньше 128:



C. 2063696 всего слов. Имеем следующую мапу по популярности:

dict.csv

Данный файл представляет из себя объединение русско-мансийских словарей mansi translator и [fu-lab](fu-lab). Словарь содержит 11506 уникальных строк.