Содержание

1 Сбор художетсвенной литературы для корпуса. Китайский

Так как в открытом доступе отсутствуют корпуса художественных текстов на китайском языке, то на данном этапе пришлось самому строить такой корпус. Текста были взяты с сайта libgen раздела художественной литературы на китайском языке, с помощью написанного специально для этой задачи парсера, который обходит дерево HTML страницы и получает ссылки на скачивания файлов в формате file.epub, file.mobi, file.txt. На выходе получается файл в формате txt, содержащий ссылки на все обнаруженные парсером текстовые файлы. Далее файл напрямую скачивается по ссылке. По итогу корпус представляет из себя файл "zh-corpus.txt"размером 6.5GB, содержащий в себе 8358 текстов на китайском языке (каждая строка это отдельный документ). Парсер был написан на языке программирования Python 3, с использованием библиотеки Selenium, позволяющий автоматизировать действия веб-браузера.

2 Сбор художественной литературы для корпуса. Бенгальский

Аналогично, в открытом доступе отсутствуют корпуса художественных текстов на бенгальском языке и поэтому пришлось строить корпус самостоятельно. Текста были взяты с сайта ebanglalibrary. Также как и в случае с libgen, парсер обходит дерево HTML страницы и получает ссылку на страницу на которой находится текст, на выходе получается файл со страницами с текстом. Далее отдельно по получаенным ссылкам, обходятся страницы и из них вычленяется содержимое в тегах <p>. По итогу получился корпус состоящий из приблизительно 30000 текстов суммарно размером 1GB.

3 Получение очищенного текстового корпуса. Общая задача

Основной задачей на данном этапе является очистка документов от:

- Служебных слов - предлоги, союзы, служебные частицы и т.д.
- Шумовых символов - пунктуация, перенос строки, не относящиеся к языку символы.
- Имен собственных и местоимений - однако чтобы не нарушалась смысловая структура предложения, слова из этой группы замещаются на абстрактное слово "объект".
- Имен числительных - ананлогично именам собственным и местоимениям заменяются на слово "число".
- Слова остальных частей речи, относящиеся к группе знаменательных необходимо лемматизировать.

Пример предобработки текста на русском языке:
Был текст: Он распахнул окно и увидел 3 птицы на дереве
Стал текст: Объект распахнуть окно увидеть число птица дерево

4 Особенности предобработки художественных текстов на китайском языке.

В китайском языке отсутствует необходимость в лемматизации, так как все слова существуют в единтвенной грамматической форме, грамматическую особенность слову придают особые служебные иероглифы-частицы. Поэтому из 3 необходимо выполнять только действия 1-4. Однако в китайском языке между словами отсутствуют разделители, поэтому необходимо уметь разбивать предложение на слова-токены. С этими задачами справляется библиотека spacy.io, использующая предобученную на блогах, новостях, комментариях модель.

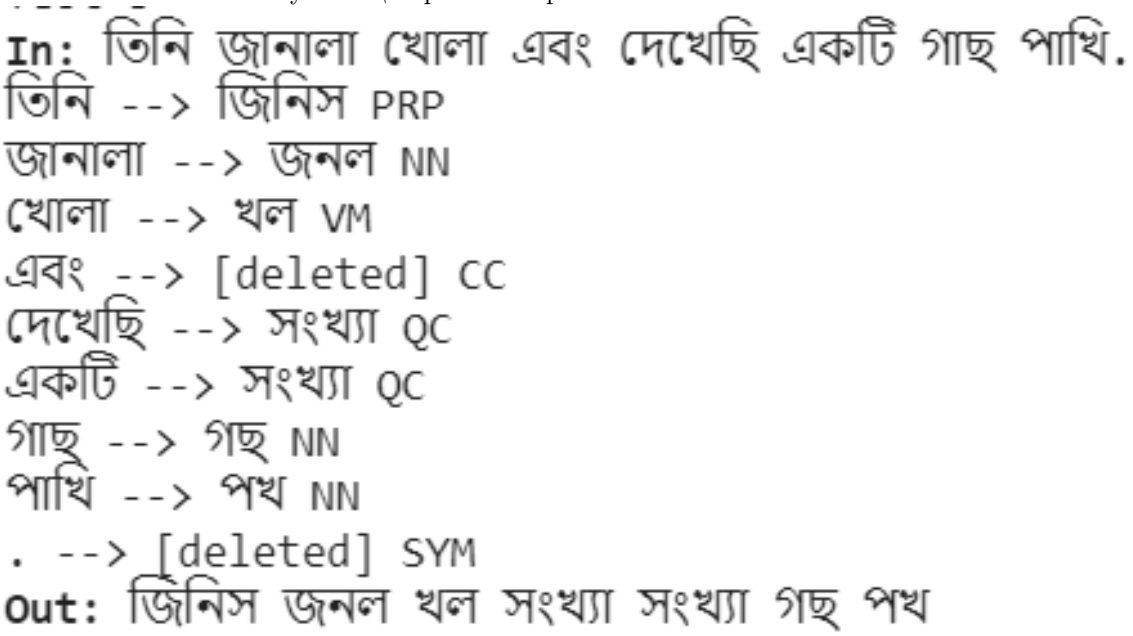
Данная языковая модель разбивает текст на токены, где для каждого токена определен его класс POS, в приложении A.2 указаны все классы POS и отмечены (×) классы, токены которых нужно удалять из текста, (◇) классы, токены которых заменять на слово " 物体"(перевод: объект) и (♠) классы, токены которых нужны замнять на слово 数字 "(перевод: число).

5 Особенности предобработки художественных текстов на бенгальском языке.

Так как в бенгальском языке присутствуют различные формы слова, то для предобработки текста необходимо выполнять все этапы из 3. Для выполнения этой задачи использовалась библиотека Spark NLP с предобученной моделью lemma-bn и pos-msri-bn для лемматизации и POS классификатора соответственно. Токенайзер разбивает текст на токены, где для каждого токена определен его класс POS, и лемма, в приложении A.1 указаны все классы POS и отмечены (×) классы, токены которых нужно удалять из текста, (⋄) классы, токены которых заменять на слово "(перевод: объект), (♠) классы, токены которых нужны замнять на слово - (перевод: число) и (♣) классы, токены которых нужно лемматизировать.

Ниже получен пример работы алгоритмы предобработки текста, вывод вида "токен –> лемма POS":

Рис. 1: Визуализация работы обработчика текста на бенгальском языке



5.0.1 TF-IDF матрица

Пусть D - некоторая коллекция из m текстов (корпус), $d \in D$ - некоторый текст из этого корпуса, всего уникальных слов-токенов (далее просто "слово") в корпусе n, $w \in d$ - слово, содержащееся в тексте d, n_w - число вхождений слова w в тексте, тогда $TFIDF(w, d)$ term frequency — inverse document frequency- статистическая мера, отражающая релевантность слова w в тексте d относительно корпуса D, которому текст d принадлежит. Вычисляется по следующим образом:

$$TFIDF(w, d) = TF(w, d) \times IDF(w, D) \tag{1}$$

Где

$$TF(w, d) = \frac{n_w}{\sum_{w' \in d} n_{w'}} \tag{2}$$

$$IDF(w, d) = \log \frac{|D|}{|\{d_i \in D \mid w \in d_i\}|} \tag{3}$$

где $|D|$ размер корпуса, $|\{d_i \in D \mid w \in d_i\}|$ число текстов, содержащих слово w. Получаем матрицу A, где:

$$a_{ij} = TFIDF(w_j, d_i) \tag{4}$$

5.0.2 SVD разложение

К полученной матрице A применяется SVD разложение ранга r [1]:

$$A \approx \hat{A} = U\Sigma V^T \tag{5}$$

где $U \in Mat(\mathbb{R})_{m\times r}$ и $V \in Mat(\mathbb{R})_{n\times r}$ ортогональные матрицы, $\Sigma \in Mat(\mathbb{R})_{r\times r}$ со значениями на диагонали $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Получаем итоговое представление слова в словаре, в котором:

$$w_i \mapsto v_i \cdot \Sigma \tag{6}$$

где v_i i-й столбец матрцицы V.

Список литературы

[1] Jerome R. Bellegarda. *Latent Semantic Mapping: Principles and Applications*. SYNTHESIS LECTURES ON SPEECH AND AUDIO PROCESSING. Morgan and Claypool.

6 Приложения

A Расшифровка POS-тэгов, выделяемых используемыми языковыми моделями.

A.1 Бенгальский язык. POS классификатор pos-msri-bn для Spark NLP

- NN - noun (♣)
- SYM - symbol (×)
- NNP - propper noun (⋄)
- VM - modal verb (♣)
- INTF - intesifier (×)
- JJ - Adjective (♣)
- QF - Quantifiers (×)
- CC - coordinating conjunction (×)
- NST - noun (♣)
- PSP - adposition (×)
- DEM - pronoun (⋄)
- PRP - posessive pronoun (⋄)
- NEG - negative (×)
- WQ - wh-qual (×)
- RB - adverb (♣)
- VAUX - Verb Auxiliary (×)
- UT (×)
- XC (×)
- RP - particle (×)

- Q0 - ordinal number (♠)
- QC - cardinal number (♠)
- BM - (×)
- NNC - compound noun (♣)
- PPR - postposition (×)
- INJ - (×)
- CL - (×)
- UNK - (×)

A.2 Китайский язык. POS классификатор zh-core-web-lg

- ADJ - adjective (♣)
- ADP - adposition (×)
- ADV - adverb (×)
- AUX - auxilary (×)
- CONJ - conjunction (×)
- CCONJ - coordinating conjunction (×)
- DET - determiner (×)
- INTJ - interjection (×)
- NOUN - noun (♣)
- NUM - numeral (♠)
- PART - particle (×)
- PRON - pronoun (◇)
- PROPN - proper noun (◇)
- PUNCT - punctuation (×)
- SCONJ - suboridnating conjunction (×)
- SYM - symbol (×)
- VERB - verb (♣)
- X - other (×)
- SPACE - space (×)