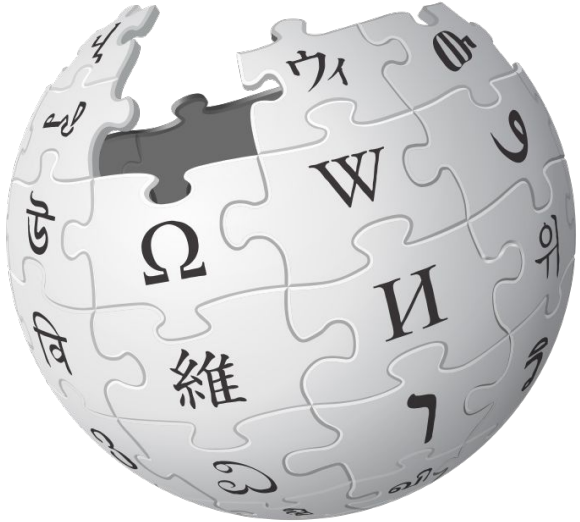
Распознавание языка текста

Презентацию подготовил
студент 1 курса СПбГУ
Леонов Даниил

Постановка задачи

Необходимо написать классификатор текстов по языку на основе данных википедии



WIKIPEDIA
The Free Encyclopedia

Данные

Какие были варианты?

- Готовые датасеты
  - Wili-2018 (по 1000 документов на 235 языках)
  - Language-detection (10267 документов на 17 языках)
  - papluca/language-identification (90 000 документов на 20 языках)
- Дампы википедии в формате zim
- Сделать своими руками

Создание датасета

- Европейские языки индо-европейской группы
- Только латиница и кириллица (то есть без греческого)
- Одинаковое количество документов для всех языков
- Документы получаются случайным образом (https://lang.wikipedia.org/wiki/Special:Random)
- 275000 документов на 23 языках

Языки

Wiki Code	Language				
be	Belarusian	fr	French	sk	Slovak
bs	Bosnian	ga	Irish	sl	Slovene
bg	Bulgarian	mk	Macedonian	es	Spanish
cs	Czech	nl	Dutch	sq	Albanian
cy	Welsh	pl	Polish	sr	Serbian
da	Danish	pt	Portuguese	sv	Swedish
de	German	ro	Romanian	uk	Ukrainian
en	English	ru	Russian		

Распознавание

Векторизация

- CountVectorizer
- TfidfVectorizer
- Doc2Vec

Модели

- MultinomialNB
- RandomForestClassifier
- LinearSVC
- LogisticRegression

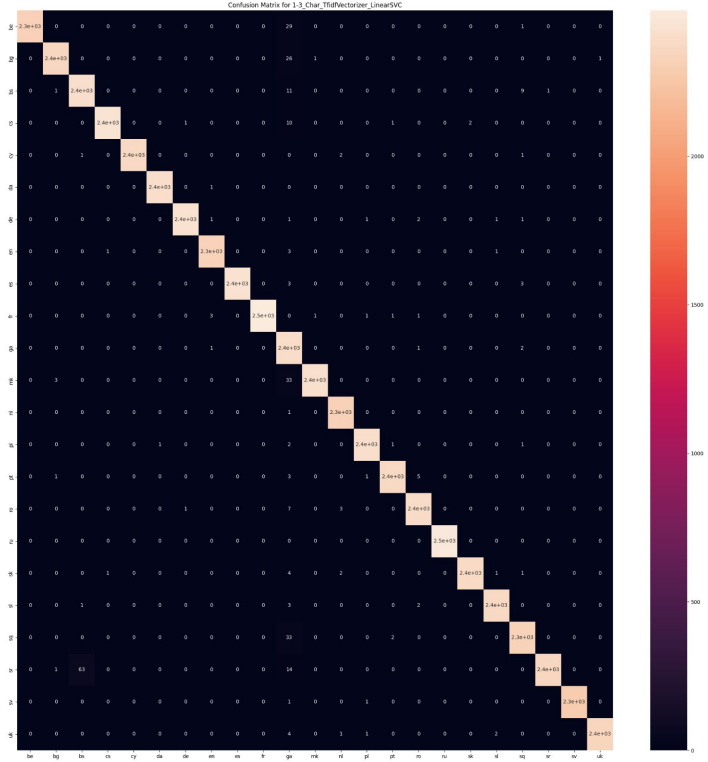
Результаты

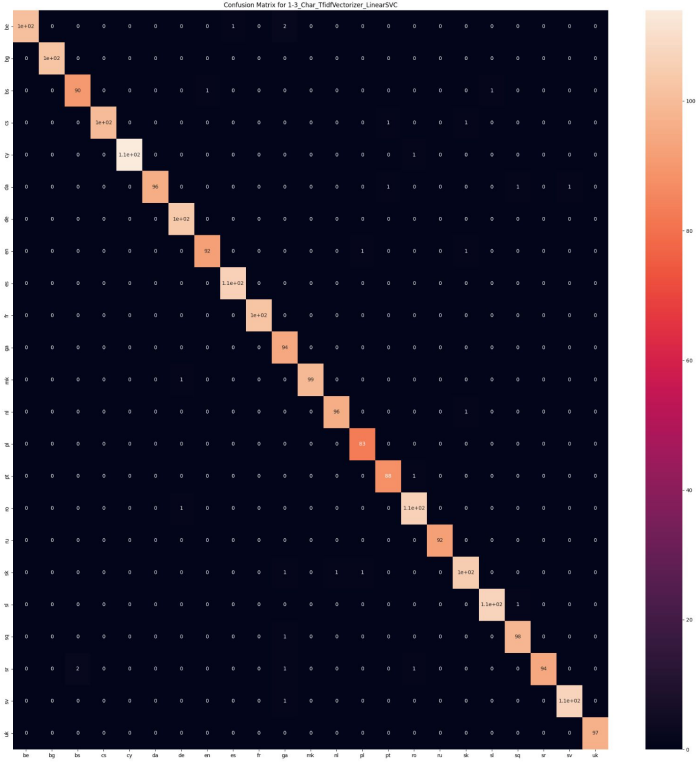
На тестовом сплите (20%)

	1-1_MultinomialNB	1-1_RandomForestClassifier	1-1_LinearSVC	1-1_LogisticRegression	1-3_MultinomialNB	1-3_RandomForestClassifier	1-3_LinearSVC	1-3_LogisticRegression
accuracy	0,9859	0,9935	0,9940	0,9927	0,9859	0,9933	0,9940	0,9927
precision	0,9865	0,9938	0,9943	0,9930	0,9865	0,9937	0,9943	0,9930
recall	0,9860	0,9935	0,9941	0,9927	0,9860	0,9934	0,9941	0,9927
f1	0,9860	0,9936	0,9941	0,9928	0,9860	0,9934	0,9941	0,9928

На наборе из 4600 слов (тест на 50%)

	1-1_MultinomiaINB	1-1_RandomForestClassifier	1-1_LinearSVC	1-1_LogisticRegression	1-3_MultinomiaINB	1-3_RandomForestClassifier	1-3_LinearSVC	1-3_LogisticRegression
accuracy	0,8970	0,9800	0,9883	0,9822	0,8970	0,9778	0,9883	0,9822
precision	0,9244	0,9810	0,9882	0,9823	0,9244	0,9794	0,9882	0,9823
recall	0,9021	0,9802	0,9883	0,9824	0,9021	0,9782	0,9883	0,9824
f1	0,8918	0,9801	0,9882	0,9822	0,8918	0,9781	0,9882	0,9822


Confusion Matrix for 1-3_Char_TfidfVectorizer_LinearSVC
be bg bs cs cy da de en es fr ga mk nl pl pt ro ru sk sl sq sr sv uk
2000
1500
1000
500
0


Confusion Matrix for 1-3_Char_TfidfVectorizer_LinearSVC
be bg bs cs cy da de en es fr ga mk nl pl pt ro ru sk sl sq sr sv uk
100
80
60
40
20
0

Doc2Vec

	RandomForestClassifier	LinearSVC	LogisticRegression
accuracy	0,8060869565	0,7973913043	0,5965217391
precision	0,8087557178	0,812629275	0,6589722188
recall	0,8084228487	0,8031183974	0,6016980823
f1	0,8038107109	0,7826804129	0,595524339

Выводы

Что использовать?

- Даже на 1/60 датасета работает хорошо
- Смена размера n-грамм tfidf практически не влияет на результат
- Лучшие результаты показывает LinearSVC