Real-World Applications: TF-IDF

In this task Hadoop Streaming is used to process Wikipedia articles dump (/data/wiki/en_articles_part).

The purpose of this task is to calculate tf*idf for each pair (word, article) from the Wikipedia dump. Apply the stop words filter to speed up calculations. Term frequency (tf) is a function depending on a term (word) and a document (article):

tf(term, doc_id) = Nt/N,

where Nt - quantity of particular term in the document, N - the total number of terms in the document (without stop words)

Inverse document frequency (idf) is a function depends on a term:

idf(term) = 1/log(1 + Dt),

where Dt - number of documents in the dataset with the particular term.

You can find more information here: https://en.wikipedia.xn--org/wiki/Tfidf-q82h but use just the formulas mentioned above.

Dataset location: /data/wiki/en_articles_part

Stop words list is in *'/datasets/stop_words_en.txt'* file.

Format: article_id <tab> article_text

To parse the articles don't forget about Unicode (even though this is an English Wikipedia dump, there are many characters from other languages), remove punctuation marks and transform words to lowercase to get the correct quantities. To cope with Unicode we recommend to use the following tokenizer:

Output: tf*idf for term='labor' and article_id=12

The result on the sample dataset:

```
0.000351
```

Hint: all Wikipedia article_ids are greater than 0. So you can use a dummy article_id=0 to calculate the number of documents with each term.

If you want to deploy the environment on your own machine, please use bigdatateam/yarn-notebook Docker container.

Для оптимизации учебного процесса в этом курсе используется сторонний инструмент Real-World Applications: TF-IDF. Инструмент получает основные сведения: идентификатор Coursera.

□ Я, **Igor Storozhev**, понимаю, что отправка работы, выполненной посторонним лицом, может привести к недоступности этого курса или отключению моего аккаунта Coursera.
Узнайте больше о Кодексе чести Coursera

Открыть инструмент