



Big Data

Dashboard ► Mineração de Dados Complexos ► 2018 ► Venturus ► INF-0617-Venturus ►
Avaliações ► Trabalho 4: Complexidade de Vocabulário

Trabalho 4: Complexidade de Vocabulário

Introdução

O dicionário Oxford da língua inglesa lista mais de 170 mil palavras em uso atualmente, além de cerca de 47 mil palavras em desuso. Estima-se, entretanto, que um vocabulário de cerca de 3000 palavras é suficiente para compreender mais de 90% das conversas, livros, filmes, e músicas em Inglês.

Objetivo

Neste trabalho, vamos analisar a complexidade do vocabulário do subconjunto da Biblioteca Gutenberg com o qual trabalhamos neste módulo, de forma a determinar qual autor, dentre os incluídos no pacote de livros, usa o vocabulário mais complexo.

Parte I

Vamos determinar inicialmente uma lista das 3000 palavras mais utilizadas entre todos os livros do pacote. Devem ser ignorados todos sinais de pontuação, números, e as palavras devem ser todas apresentadas em letras minúsculas. O resultado desta parte do trabalho deve ser armazenado como um arquivo com 3000 linhas, cada uma contendo uma única palavra.

Parte II

Com a lista de palavras comuns, na segunda parte do trabalho analisaremos os trabalhos de cada autor para determinar sua complexidade. Para tanto, listaremos todas as palavras na obra de cada autor, sem repetições.

Em seguida, totalizaremos uma pontuação de complexidade para cada autor. Para cada palavra “comum” (ou seja, presente na lista de palavras mais utilizadas) na obra do autor, atribuiremos uma pontuação de 0, e para cada palavra “exótica” (ou seja, que não está na lista de palavras mais utilizadas), atribuiremos uma pontuação de 1. A complexidade do autor será definida como a pontuação total dividida pelo número de palavras utilizadas pelo autor.

Exemplo

Considerando o autor Donald Trump, e sua obra (tweet):

“Despite the constant negative press covfefe.”

Poderíamos ter no exemplo seis palavras no total, com cinco palavras comuns, e uma palavra exótica. A complexidade do autor seria, portanto $1/6 = 0.16$.

Entrega

Para a entrega, você deve fazer upload de um único arquivo .zip contendo o seu código, instruções sobre como executá-lo no container Docker, e um breve relatório em formato PDF ou TXT descrevendo a sua implementação e resultados.

O resultado final do trabalho deve ser um arquivo com N linhas, onde N é o número de diferentes autores na biblioteca. Cada linha deve conter o nome do autor e a sua complexidade calculada. Os resultados devem ser ordenados no arquivo da maior para menor complexidade. Como exemplo, uma linha do arquivo deve ter o formato:

```
Trump, 0.16
```

Observações:

- O arquivo “master_list.csv” contém os nomes dos autores para cada um dos arquivos na biblioteca. Note que todos os nomes de arquivos no pacote estão prefixados com u-
 - Você pode considerar apenas o sobrenome do autor para o agrupamento.
- A sua solução deve ser estruturada com alguma ferramenta do pacote Hadoop/Spark, incluindo:
 - MapReduce com Hadoop Streaming
 - PySpark
 - Pig
 - Hive
- Você pode usar combinações das ferramentas acima para resolver partes diferentes do problema.
- A sua solução não pode usar exclusivamente processamento local, mesmo que paralelo. Você deve necessariamente usar alguma das ferramentas do pacote Hadoop/Spark.

Crédito Extra

- Para receber crédito extra (10% da nota total), crie um filtro que ignore as obras que não estão em inglês, e as obras de autores desconhecidos (anonymous).

Submission status

Submission status	No attempt
Grading status	Not graded
Due date	domingo, 28 outubro 2018, 11:55
Time remaining	8 days
Last modified	sábado, 20 outubro 2018, 10:14
Submission comments	► Comments (0)

Add submission

Make changes to your submission

NAVIGATION



Dashboard

■ Site home

Site pages

Current course

INF-0617-Venturus

Participants

Badges

Apresentações

Tutoriais


Aulas

Avaliações

 Notas e Frequências

 Trabalho 1: Processamento Paralelo Básico

 Trabalho 2: MapReduce Local

 Trabalho 3: Hadoop Streaming

 **Trabalho 4: Complexidade de Vocabulário**

My courses

ADMINISTRATION



Course administration

You are logged in as Yakov Nae (Log out)
INF-0617-Venturus