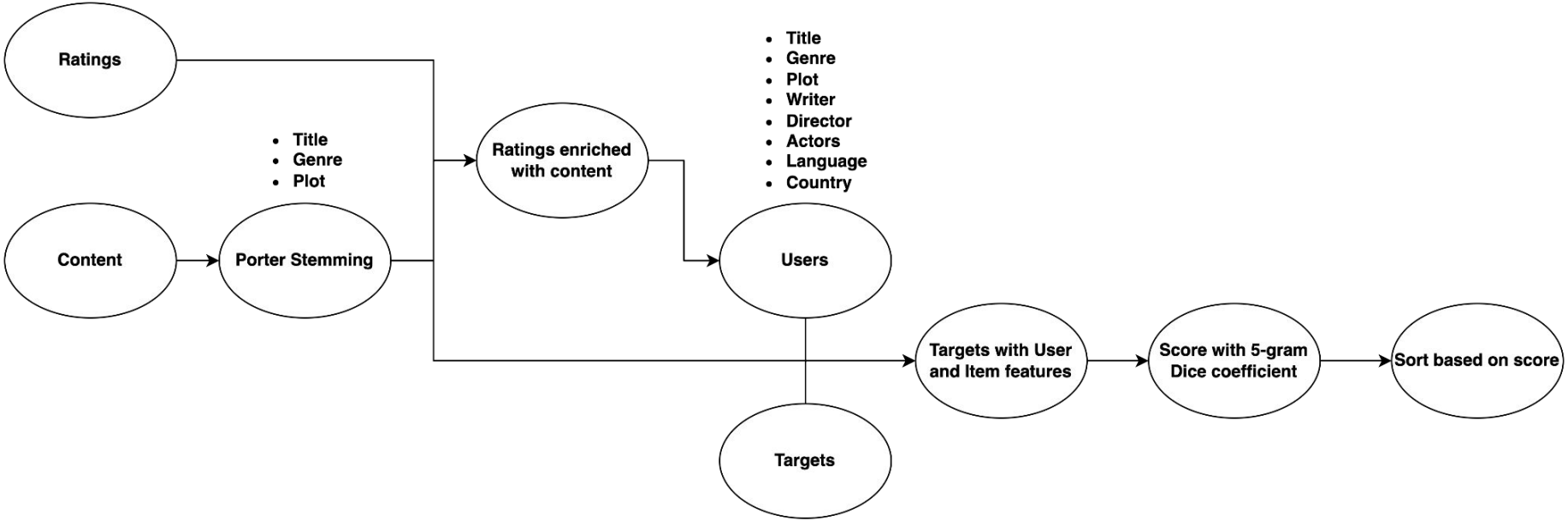
DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Research Challenge 2

DCC049 - Tópicos em Sistemas de Informação
Sistemas de Recomendação

Danilo Pimentel de Carvalho Costa

Content-based Recommendation



Observações

- Problema de cold-start para itens é resolvido por content-based recommendation
- Performance aumenta utilizando mais features textuais:
  - Title, Plot, Genre: 0.23327
  - Title, Plot, Genre, Writer, Directors, Actors, Language, Country: 0.35362
- Performance aumenta utilizando n-grams maiores:
  - 2-gram: 0.35362
  - 3-gram: 0.37050
  - 4-gram: 0.37756
  - 5-gram: 0.38050

Possíveis otimizações

- Uso de TF-IDF nos campos textuais para detectar relevância de termos
- Uso de *ratings* para determinar peso para itens na representação de usuário, ou somente utilizar itens com avaliação maior que X.
- Uso de collaborative-filtering como outro componente para o score
  - Agregação de ratings de vizinhos pode ser usado como peso para score do recomendador baseado em conteúdo
- Uso de outras features
  - imdbRating: filmes populares vs. filmes controversos
  - imdbVotes: filmes *mainstream* vs. filmes alternativos
  - Year: filmes clássicos vs. filmes recentes

Amazon Sagemaker

Jupyter Notebooks com máquinas de propósito geral, otimizadas para memória, computação ou aceleração por GPU

Referências

https://tartarus.org/martin/PorterStemmer/

https://github.com/luozhouyang/python-string-similarity

https://www.datacamp.com/community/tutorials/stemming-lemmatization-python