Representação de Elementos de As Crônicas de Gelo e Fogo Utilizando NLP

Nélio Cezar Muniz Sampaio
Orientador: Adriano Alonso Veloso

nelio.sampaio@dcc.ufmg.br

Resumo

Neste trabalho, propõe-se a implementação de um modelo baseado em processamento de linguagem natural que permita a extração e representação dos diversos objetos presentes na série de livros A Song of Ice and Fire de George R R Martin. Propõe-se ainda que tal modelo possa ser representado de forma gráfica coerente com o que é apresentado na história.

Introdução

Em aprendizado de máquina, as diversas redes neuronais são formadas por neurônios, que são as unidades responsáveis por aprender características dos dados e o comportamento esperado dada uma entrada. No aprendizado supervisionado, o neurônio recebe uma entrada e a resposta esperada para o dado, então ele gera uma resposta própria e compara com a correta.

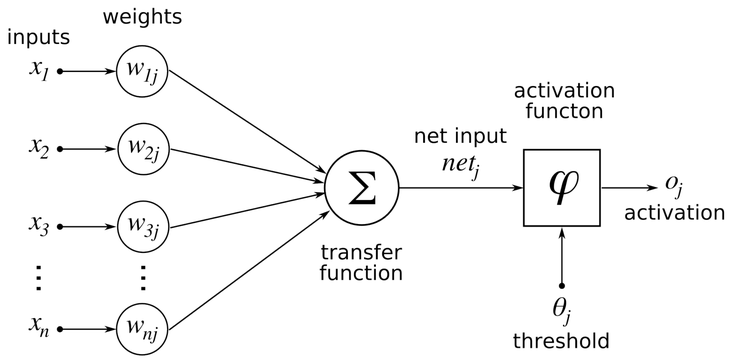

Figura 1 – Neurônio

Agrupando os neurônios de formas específicas, são criadas redes capazes de aprender comportamentos complexos de acordo com os dados fornecidos. Em Processamento de Linguagem Natural (Natural Language Processing – NLP) é muito importante a etapa de extração do significado das palavras.
O algoritmo Skipgram gera uma representação vetorial para cada palavra presente no texto, no caso os livros da série. Essa representação é chamada de embedding.

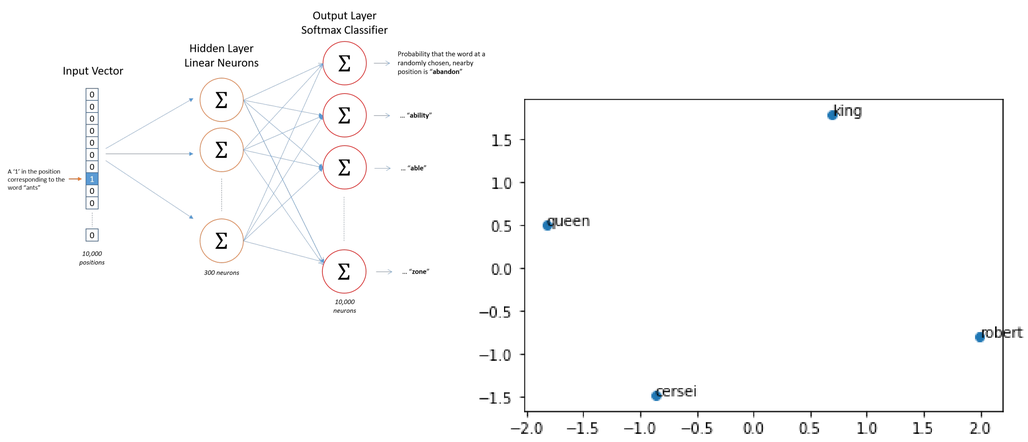

Figura 2 – Representação do Skipgram e resultados obtidos

Método

Para o desenvolvimento deste trabalho, foram utilizadas as versões em inglês dos livros da série “A Song of Ice and Fire” de George R R Martin. Os livros foram reunidos e tratados como um texto único, foram removidas stop words e feito o preprocessamento de forma a manter nomes compostos como castle black para que pudessem ser utilizados corretamente durante a análise. Foi gerado o modelo apresentado utilizando o skipgram e os gráficos apresentados utilizaram o PCA para redução de dimensionalidade dos embeddings.

Resultados

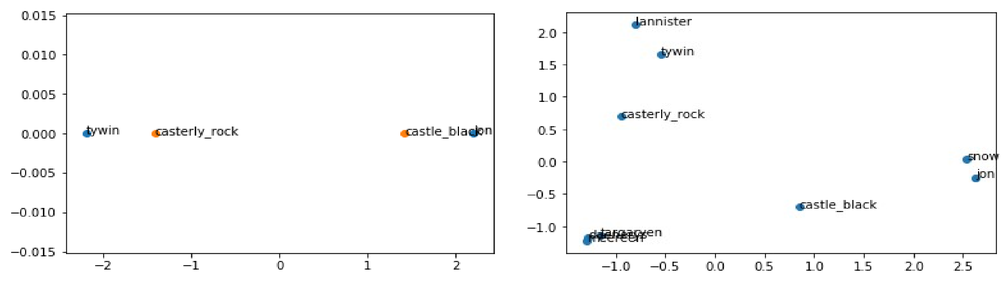

Figura 3 – Relações entre personagens, castelos e nomes

Os gráficos gerados refletem em sua maioria configurações que coerentes com o que é apresentado ao longo da série. Nos casos da figura 3, é possível perceber um agrupamento entre os pontos de elementos que compõem um contexto similar em comparação com os outros.

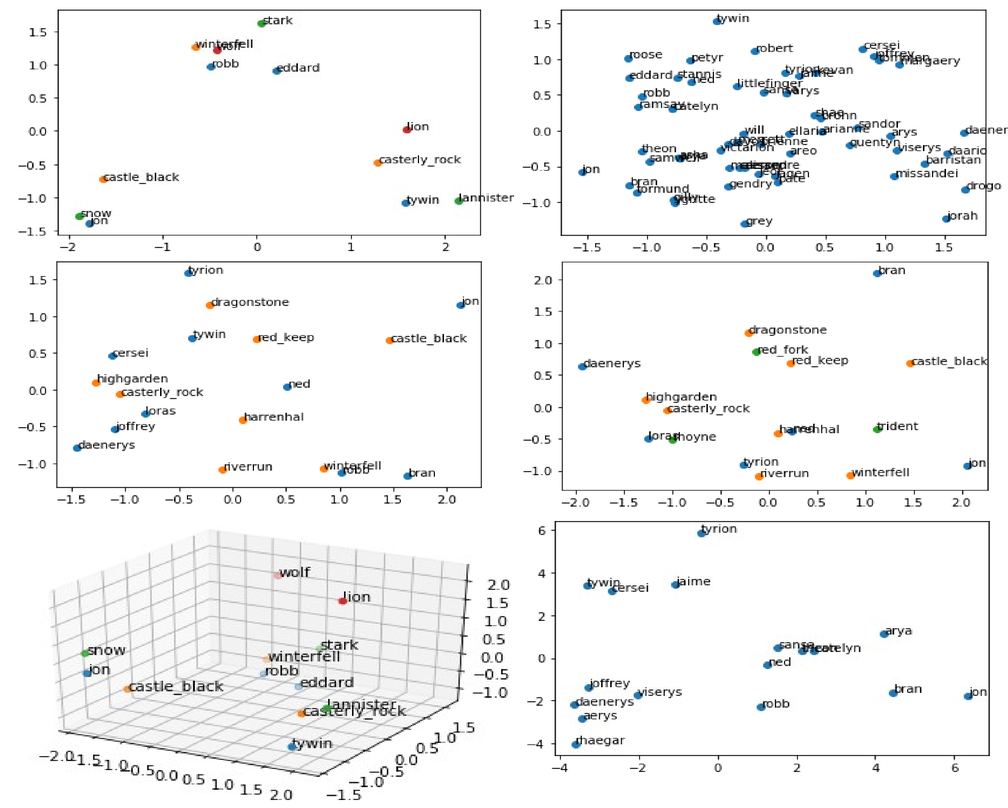

Figura 4 – Resultados mostrando relações entre outros tipos de embeddings

Conclusão

Neste projeto, foi possível representar alguns dos elementos do universo de As Crônicas de Gelo e Fogo utilizando word2vec As representações ilustram de forma coerente diversas das relações buscadas, mas ainda não se mostrou suficiente para representá-los ao aumentar a complexidade, ou seja, a quantidade e diversidade de elementos.
Para este projeto ainda deverão ser buscadas técnicas que permitam melhorar as representações geradas de forma a construir uma gama de representações com diversos elementos. O aumento na qualidade das representações poderá permitir sua utilização em outros casos como política e história.

Supported by

CAPES
FAPEMIG
CNPq
ICEx Instituto de Ciências Exatas UFMG

Referências

[1] Laurens van der Maaten & Geoffrey Hinton (2008). Visualizing Data Using t-SNE. Journal of Machine Learning Research.
[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean (2013). Distributed Representations of Words and Phrases and their Compositionality.
[3] Bengio Y., Schwenk H., Senécal JS., Morin F., Gauvain JL. (2006) Neural Probabilistic Language Models. In: Holmes D.E., Jain L.C. (eds) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, vol 194. Springer, Berlin, Heidelberg
[4] Página Hackernoon. Disponível em: https://hackernoon.com/word-embeddings-in-nlp-and-its-applications-fab15eaf7430. [Acessado em 08/09/2019].
[5] Introduction to t-SNE. DataCamp. Disponível em: https://www.datacamp.com/community/tutorials/introduction-t-sne. [Acessado em 08/09/2019].
[6] https://towardsdatascience.com [Acessado em 25/10/2019]
[7] https://prakhartechviz.blogspot.com [Acessado em 25/11/2019]
[8] https://matplotlib.org [Acessado em 25/11/2019]