Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

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

Aluno: Nélio Cezar Muniz Sampaio
Tipo do trabalho: Pesquisa científica
Orientador: Adriano Alonso Veloso
Relatório de projeto
Projeto Orientado em Computação 2

Belo Horizonte, 3 de Dezembro de 2019

Resumo

Neste trabalho foram utilizadas técnicas de processamento de linguagem natural em conjunto com os livros da série A Song of Ice and Fire, de George R R Martin, com intuito de representar os diversos elementos presentes na série de forma coerente. Nos experimentos realizados, algumas as representações refletem vários dos contextos presentes na história contada na série de livros. Entretanto, também foram encontrados algumas dificuldades como para representar vários elementos de tipos diversos, onde não há clareza nos grupos apresentados.

Introdução

Em 1950, Alan Turing propôs um teste para se definir se uma máquina era ou não inteligente, e desde então, o chamado Teste de Turing se tornou um ponto de partida para diversos trabalhos que hoje se enquadram na chamada inteligência artificial. O teste consiste uma pessoa A que se comunica com um pessoa B e com um computador, estando os três separados uns dos outros; ao se comunicar, a pessoa A não sabe se quem está respondendo é a pessoa B ou o computador, e o teste diz que, se não for possível distinguir a resposta da máquina da resposta de uma pessoa, a máquina passou no teste e é inteligente. O teste marca o início da inteligência artificial, e a área foi fortalecida por uma série de trabalhos publicados na época. Quase 70 anos depois, a computação tem hoje tecnologia que possibilita a criação de agentes muito superiores aos que podiam ser imaginados na época.

Natural Language Processing - NLP- ou Processamento de Linguagem Natural é o nome dado à subárea da inteligência artificial e da linguística responsável por estudar e desenvolver o processamento das diversas linguagens naturais possibilitando o entendimento e a geração pelos computadores. O desenvolvimento da área possibilitou a criação de diversas aplicações atuais, várias delas sendo exploradas comercialmente. Assistentes virtuais como Google Assistant, Cortana e Alexa são exemplos já amplamente explorados pelas empresas gigantes da tecnologia Google, Microsoft e Amazon, respectivamente. Além disso, agentes conversacionais estão cada vez mais presentes entre as empresas, técnicas de análise de sentimentos e modelos de question answering, entre diversas aplicações.

Este trabalho busca abordar NLP de forma a extrair informações não explícitas dentro do texto utilizado como base. Assim, busca-se entendimento das características descritas em uma parte e a organização em conjunto com o que é dito ao longo do texto por completo. O texto a ser utilizado é a série de livros de fantasia A Song of Ice and Fire (As Crônicas de Gelo e Fogo) de George R R Martin. A série foi escolhida, dentre outros motivos, por possuir um material extenso e contar

com um estilo de escrita bem descritivo em relação a cenários, personagens e à estrutura social.

O trabalho consiste em gerar um modelo a partir de técnicas de NLP e estatística de redução de dimensionalidade de dados para representar o conhecimento obtido a partir dos livros. A representação do modelo deve ser feita de acordo com as características selecionadas (personagens, casas nobres, cidades, castelos, etc) de forma gráfica coerente com o que é descrito na série.

Referencial Teórico

Word2vec é um conjunto de modelos utilizados para representar as palavras de um texto em forma de vetores. Neste caso, o texto utilizado como base foram os livros que compõem a série As Crônicas de Gelo e Fogo. Assim, no modelo gerado, cada uma das palavras presentes no texto possui uma representação vetorial a ser utilizada nos experimentos.

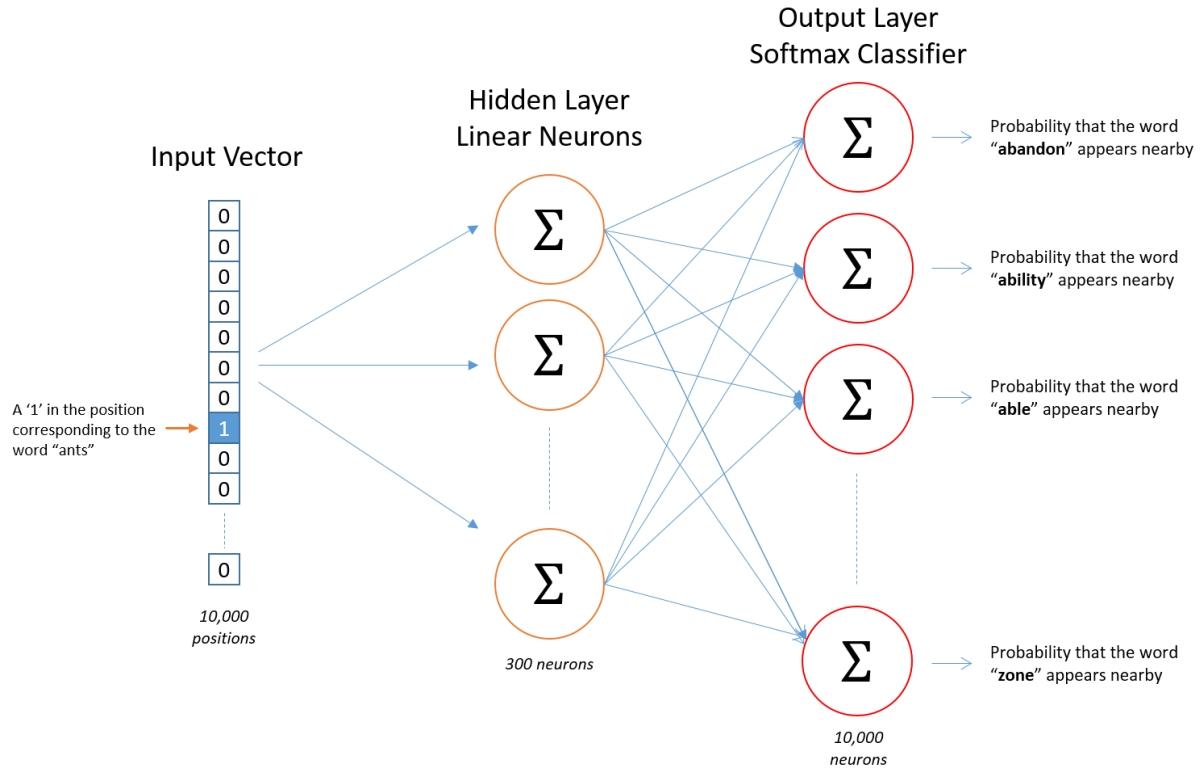

figura 1 - estrutura do algoritmo skipgram

O algoritmo escolhido para converter as palavras em vetores foi o skipgram, devido a sua representatividade dentro do processamento de linguagem natural mediante sua simplicidade. O modelo é gerado ao treinar uma rede de neurônios para que, dada uma palavra w presente no texto, seja possível prever quais palavras fazem parte do seu contexto. O contexto é dado por uma janela de tamanho definido c, e

que englobe as c palavras anteriores antes e c palavras depois de w. A palavra dada como entrada pe colocada no formato one-hot, onde cada palavra é representada por um vetor de tamanho igual ao número de palavras do texto e todas as posições tem valor 0, exceto a posição de índice igual ao índice da palavra. Dessa forma, ao treinar a rede, apenas serão atualizados pesos específicos para cada palavra, resultando em uma representação que leve em consideração o contexto em que cada palavra aparece ao longo do texto.

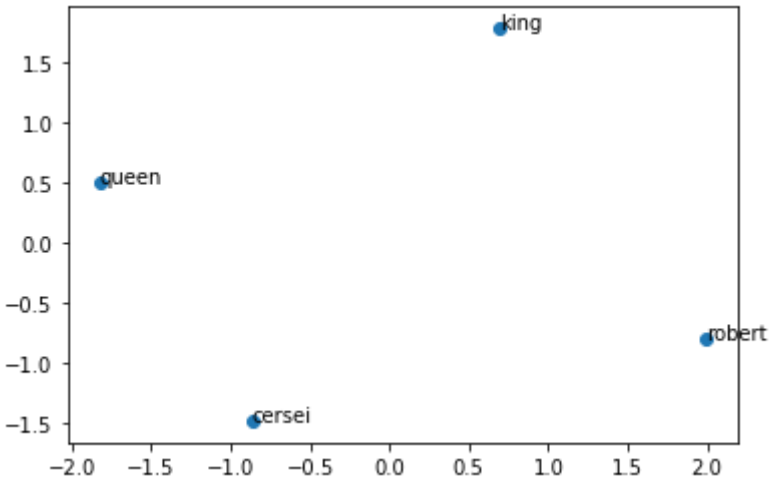

figura 2 - Representação dos embeddings para rei, rainha, cersei e robbert.

É importante ressaltar que a rede não é utilizada de fato para o que ela foi treinada, mas sim os pesos que compõem os neurônios são utilizados para representar as palavras do texto. Cada dimensão do vetor, chamado de embedding, representa características aprendidas para a palavra em questão. A imagem 2 representa em um espaço de duas dimensões os embeddings para as palavras queen, king, cersei e robert, sendo as últimas duas referentes a dois personagens presentes nos livros, uma rainha e um rei respectivamente. Com esse exemplo é possível perceber como o sentido das palavras é mantido através da similaridade da relação entre palavras de sentido equivalente.

A representação de cada embedding de forma gráfica ao longo deste trabalho utiliza o algoritmo de Principal Analysis Component (PCA) para fazer a redução de dimensionalidade de 100 para 2 ou 3.

Resultados

A partir das representações obtidas foi feita uma análise qualitativa a partir dos contextos e narrativas presentes na história contada pela série. Na figura 3 por exemplo, estão representados 3 personagens, 3 castelos e o nome de 3 casas, sendo que um de cada categoria está relacionado com um da outra formando 3 trios.

Ao representar as palavras no espaço de duas dimensões é possível perceber que elas se agrupam de acordo com a relação entre elas de acordo com a história.

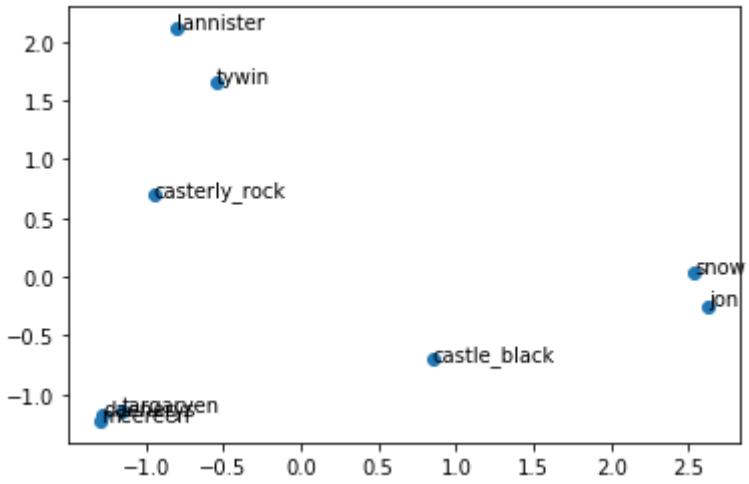

figura 3 - Embeddings de personagens, casas e castelos

E essa análise pode ser expandida para outros elementos ou mesmo um número diferente de palavras.

figura 4 - (a) representação de personagens da série; (b) personagens com castelos, símbolos e nome da casa

Na figura 4 (a) estão dispostos diversos personagens presentes na série, sendo que, ao representar tais tipos de palavras, é esperado que o posicionamento de cada uma delas fique de acordo com a proximidade que cada personagem teve durante a trama. Esse comportamento pode ser observado entre diversos personagens de vários núcleos da história, entretanto, para alguns acaba não fazendo sentido seu posicionamento. Uma das causas que podem ser apontadas como causadoras desse comportamento é a redução de dimensionalidade, onde inevitavelmente haverá perda de informação.Outra causa pode ser o tamanho da base utilizada para o treino do modelo. Aplicações de NLP no geral utilizam bases com ordens de grandeza de diferença de tamanho para o tamanho dos 5 livros utilizados como base.

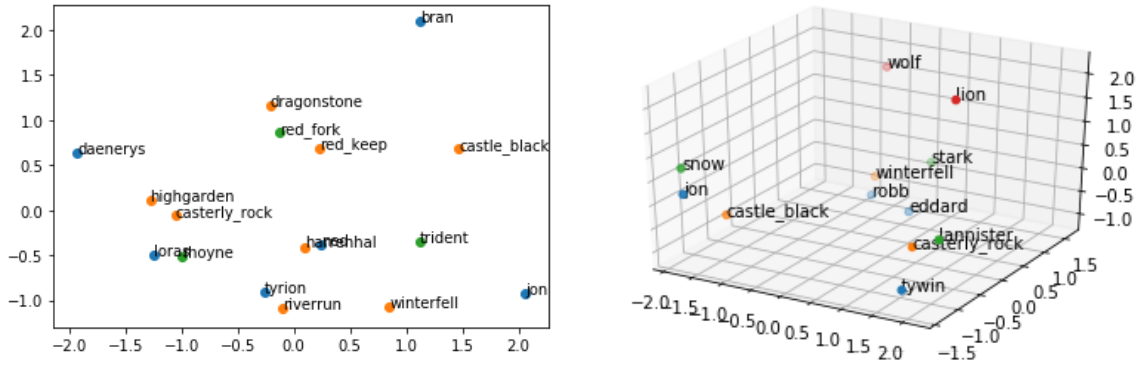

figura 5 - (a) Embeddings para personagens, castelos e rios. (b) Representação em 3 dimensões.

Com intuito de elaborar uma representação que consiga relacionar diversos elementos e de forma coerente, foram plotados dados de diferentes natureza, como a figura 5. Entretanto, nesses casos, é perceptível que a organização dos dados passa a ignorar a presença de outras palavras e se agrupam apenas com palavras de mesma natureza. Uma das tentativas de se contornar tal problema, foi aumentando a dimensão da representação. A figura 5 (b) ilustra os dados plotados em 3 dimensões. Entretanto, mesmo com o aumento de dimensões os resultados se mostraram similares, conseguindo manter um pouco de coerência entre dados de mesmo tipo, mas não fazendo sentido em relação aos demais.

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.

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]