NLP: Trabalho Prático 1 (Word2Vec)

João Mateus de Freitas Veneroso

Departamento de Ciência da Computação da Universidade Federal de Minas Gerais

September 19, 2017

Introdução

Esse relatório descreve a implementação do trabalho prático 1 da disciplina NLP. O trabalho consistiu em comparar o conteúdo de 70 obras clássicas disponibilizadas publicamente pelo projeto Gutemberg com a utilização de vetores semânticos. A versão mais recente do código pode ser obtida em:
https://github.com/jmfveneroso/word2vec_text_similarity.

Decisões de Projeto

O conteúdo de 70 obras clássicas foi o obtido no endereço:
https://www.gutenberg.org/ em formato de texto simples. Referências alheias ao conteúdo original das obras foram removidas manualmente à medida do possível para evitar contaminação dos resultados por metadados e notas dos editores. A pontuação foi removida e as letras maiúsculas foram convertidas em minúsculas. Por último, o algoritmo de *stemming* de Porter foi aplicado para extrair os radicais das palavras. Os dados tratados foram divididos em 70 arquivos contendo *tokens* referentes ao conteúdo das respectivas obras separados por espaços.

A coleção completa possui 10.593.149 palavras para um vocabulário de 22.584 termos únicos (após a realização do *stemming*). Se utilizássemos o vocabulário completo para calcular as matrizes de similaridade, elas ficariam com aproximadamente 510.037.056 entradas por documento, resultando em um espaço de aproximadamente 2.04 GB. Portanto, essa abordagem se torna rapidamente inviável em máquinas convencionais. A fim de tratar o problema, apenas os 1000 termos com maior frequência foram utilizados, excluindo-se as *stop words* mais comuns da língua inglesa.

Essa abordagem tende a subestimar a dissimilaridade entre as obras, pois ela ignora a maior parte dos termos únicos ou pouco frequentes na coleção, no entanto, esses termos são provavelmente os identificadores mais icônicos de

um autor ou obra. Por outro lado, o enfoque nas palavras de uso corriqueiro possibilita um cálculo mais robusto dos vetores semânticos, uma vez que elas aparecem em um número maior de contextos distintos na coleção.

Os vetores semânticos foram calculados com base na implementação canônica do Word2Vec. O modelo utilizado foi o *Skip Gram* com janela de tamanho 10 e vetores de 100 dimensões. Adicionalmente, foi realizado um teste com vetores de 200 dimensões. Em ambos os casos, uma matriz de similaridade quadrada de tamanho 1000 X 1000 com as distâncias por cosseno dos vetores semânticos foi computada para cada uma das obras e elas foram comparadas por meio da norma de Frobenius, que pode ser definida pela expressão:

$$\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij}-b_{ij})^2}$$

A norma de Frobenius é menor quanto mais similares forem as matrizes. No caso desse trabalho, a norma tende a ser menor quando duas obras possuem conteúdos similares ou, mais especificamente, quando os vetores semânticos são similares. A norma de Frobenius fica com o valor zero para conteúdos idênticos.

Resultados

As matrizes das 70 obras foram comparadas uma a uma por meio da norma de Frobenius resultando em um total de 4.900 combinações. Os resultados completos estão disponíveis no endereço:
https://docs.google.com/spreadsheets/d/1DvfqgbsrPrO40C6j0EnztfBqnmAAlRcKunEkCjL2-oI/edit?usp=sharing.

De acordo com a tabela 1 podemos observar que os dois textos mais similares são *War and Peace* de Leo Tolstoy, publicado em 1869, e *Les Miserábles* de Victor Hugo, publicado em 1862, com uma norma de Frobenius igual à 165,14. Ambas são obras realistas tardias do final do século XIX e provavelmente fazem um uso similar da linguagem, pelo menos nas traduções para a língua inglesa. Também podemos perceber que obras com o mesmo autor tendem a apresentar alta similaridade como *The Brothers Karamazov* e *Anna Karenina* de Tolstoy, *Alice's Adventures in Wonderland* e *Through the Looking-Glass* de Lewis Carrol e *Oliver Twist* e *Great Expectations* de Charles Dickens.

O romance *Wuthering Heights* foi o único a aparecer 13 vezes como a obra mais similar entre às 70 obras analisadas. Essas 13 obras são: *Gulliver's Travels*, *The Hound of the Baskervilles*, *Three Men in a Boat*, *Treasure Island*, *A Christmas Carol in Prose*, *Anne of Green Gables, Around the World in Eighty Days, The Scarlet Letter, On The Duty Of Civil Disobedience, The Picture of Dorian Gray, Peter Pan, The Secret Adversary* e *The War of the Worlds.* A razão para isso não é facilmente explicada, pois apesar de *Wuthering Heights* ser uma tragédia icônica de meados do século XIX, ela não possui muitos *features* em comum com as 13 obras citadas, com exceção talvez de *The Scarlet Letter.* Uma característica em comum dessas 13 obras, no entanto, é que em sua maioria

são obras de fantasia voltadas para um público infanto-juvenil. Talvez o estilo de escrita utilizado por Emily Brontë tenha algo em comum com essas obras, mas isso é improvável uma vez que *Wuthering Heights* é considerado um dos maiores romances da língua inglesa e também é reconhecido pela sua complexidade. Esse fato talvez revele uma possível falha nas premissas do algoritmo ou mesmo uma falha de implementação, no entanto, os coeficientes de similaridade para a maior parte dos outros casos são facilmente explicáveis e são coerentes considerando o conteúdo das obras selecionadas.

Uma outra relação de similaridade interessante foi encontrada entre *The King James Version of the Bible* e *The Complete Works of Shakespeare*, que possuem uma norma de Frobenius de 182,66. A razão para esse grau alto de similaridade provavelmente é a proximidade das data de publicação das obras: 1611 no caso de *The King James Version of the Bible* e 1589-1613 para as obras de Shakespeare; e também o país de origem: a Inglaterra. Dados esses fatores, é compreensível que as obras façam uma utilização similar da língua inglesa. E, de fato, observamos uma linguagem rebuscada significativamente diferente das outras obras presentes na coleção estudada.

Já a obra mais dissimilar entre todas as obras estudadas foi *Songs of Innocence, and Songs of Experience* de William Blake tendo sido a obra mais dissimilar em 51 das 70 comparações. Provavelmente, a razão para isso é que a obra em questão é uma coletânea de poemas que apresenta uma estrutura significativamente diferente em relação às demais obras escritas em prosa.

As tabelas 3 e 4 mostram a matriz de similaridade para as 10 primeiras obras da coleção e a tabela 2 mostra o nome e o autor das respectivas obras. Analisando o conteúdo das matrizes, podemos perceber que a mudança de dimensionalidade dos vetores não afetou significativamente as relações de similaridade. Portanto, vetores com 100 dimensões parecem ser suficientes para modelas as interações semânticas complexas nessa pequena coleção de documentos.

Conclusão

Esse relatório descreveu a implementação do trabalho prático 1 da disciplina NLP. Os experimentos realizados calcularam matrizes de similaridade para 70 obras clássicas com a utilização de vetores semânticos de 100 e 200 dimensões. O algoritmo mostrou ter uma boa capacidade de identificar similaridades entre as obras, no entanto, alguns casos, como o livro Wuthering Heights, apresentaram resultados insatisfatórios. De forma geral, os vetores de 200 dimensões não mostraram grande variação em relação aos resultados calculados com vetores de 100 dimensões.

Obra	Obra mais similar	Similaridade
A Study in Scarlet	The Adventures of Sherlock Holmes	387.81
Adventures of Huckleberry Finn	Moby Dick	450.78
Anna Karenina	The Brothers Karamazov	202.09
Wealth of Nations	David Copperfield	311.37
Emma	Great Expectations	313.90
The Adventures of Sherlock Holmes	The Return of Sherlock Holmes	268.01
Gulliver's Travels	Wuthering Heights	321.62
Crime and Punishment	Jane Eyre - An Autobiography	254.64
Autobiography of Benjamin Franklin	The Adventures of Sherlock Holmes	389.67
Pride and Prejudice	Sense and Sensibility	301.96
The Hound of the Baskervilles	Wuthering Heights	366.76
The Adventures of Tom Sawyer	On The Duty Of Civil Disobedience	334.46
Les Misérables	War and Peace	165.14
Beyond Good and Evil	The Works of Edgar Allan Poe — 2	440.72
Sense and Sensibility	Pride and Prejudice	301.96
The Return of Sherlock Holmes	The Adventures of Sherlock Holmes	268.01
The Jungle Book	A Tale of Two Cities	494.11
The Brothers Karamazov	Anna Karenina	202.09
Il Principe	A Tale of Two Cities	485.22
The Sign of the Four	The Hound of the Baskervilles	397.01
Three Men in a Boat	Wuthering Heights	343.43
The Count of Monte Cristo	Les Misérables	170.57
Second Treatise of Government	Jane Eyre - An Autobiography	497.94
Treasure Island	Wuthering Heights	374.28
War and Peace	Les Misérables	165.14
Leviathan	Ulysses	347.68
Ulysses	David Copperfield	252.14
The Life and Adventures of Robinson Crusoe	Moby Dick	344.68
The Republic	Jane Eyre - An Autobiography	294.32
Utopia	Gulliver's Travels	443.41
A Doll's House: a play	Oliver Twist	515.30
A Tale of Two Cities	Oliver Twist	256.20
A Christmas Carol in Prose	Wuthering Heights	449.12
Common Sense	Leviathan	508.89
Frankenstein	The Works of Edgar Allan Poe — 2	334.76
Pygmalion	The Adventures of Sherlock Holmes	465.84
Anne of Green Gables	Wuthering Heights	314.64
Alice's Adventures in Wonderland	Through the Looking-Glass	487.63
Don Quixote	The Count of Monte Cristo	184.74
Moby Dick	Great Expectations	257.38
The Importance of Being Earnest	Oliver Twist	525.55
David Copperfield	Anna Karenina	205.36
Around the World in Eighty Days	Wuthering Heights	360.48
Paradise Lost	Moby Dick	416.17
The Scarlet Letter	Wuthering Heights	341.11
Great Expectations	Oliver Twist	242.94
Candide	Frankenstein	423.64
Songs of Innocence, and Songs of Experience	The Complete Works of Shakespeare	396.47
The Works of Edgar Allan Poe — 1	On The Duty Of Civil Disobedience	300.24
Jane Eyre - An Autobiography	Great Expectations	246.67
Dracula	Great Expectations	246.40
The Complete Works of Shakespeare	War and Peace	167.77
The Works of Edgar Allan Poe — 2	On The Duty Of Civil Disobedience	307.53
Oliver Twist	Great Expectations	242.94
Metamorphosis	Don Quixote	519.39
The Divine Comedy	Paradise Lost	495.44
On The Duty Of Civil Disobedience	Wuthering Heights	295.70
The Picture of Dorian Gray	Wuthering Heights	334.10
Peter Pan	Wuthering Heights	428.71
The Iliad	Jane Eyre - An Autobiography	285.16
The Secret Adversary	Wuthering Heights	316.07
The Mysterious Affair at Styles	The Secret Adversary	390.30
The King James Version of the Bible	The Complete Works of Shakespeare	182.66
Wuthering Heights	The Return of Sherlock Holmes	291.84
The Strange Case of Dr. Jekyll and Mr Hyde	Treasure Island	468.26
The Tragedy of Romeo and Juliet	Ulysses	513.75
The Time Machine	Treasure Island	486.86
The War of the Worlds	Wuthering Heights	401.09
The Wonderful Wizard of Oz	Oliver Twist	507.84
Through the Looking-Glass	Alice's Adventures in Wonderland	487.63

Table 1: Obras mais similares com vetores de 100 dimensões

Índice	Obra	Autor(a)
1	A Study in Scarlet	Arthur Conan Doyle
2	Adventures of Huckleberry Finn	Mark Twain
3	Anna Karenina	Leo Tolstoy
4	Wealth of Nations	Adam Smith
5	Emma	Jane Austen
6	The Adventures of Sherlock Holmes	Arthur Conan Doyle
7	Gulliver's Travels	Jonathan Swift
8	Crime and Punishment	Fyodor Dostoyevsky
9	Autobiography of Benjamin Franklin	Benjamin Franklin
10	Pride and Prejudice	Jane Austen

Table 2: Primeiras 10 obras da coleção

	1	2	3	4	5	6	7	8	9	10
1	0.00	565.10	606.58	563.03	450.10	387.81	430.16	498.56	464.97	441.18
2	565.10	0.00	491.42	504.86	490.99	484.41	503.35	464.59	569.54	521.14
3	606.58	491.42	0.00	316.25	424.03	481.15	495.59	275.92	582.96	487.70
4	563.03	504.86	316.25	0.00	414.10	463.75	443.60	337.29	538.33	461.32
5	450.10	490.99	424.03	414.10	0.00	349.05	362.39	350.04	436.80	326.17
6	387.81	484.41	481.15	463.75	349.05	0.00	346.89	379.61	389.67	349.61
7	430.16	503.35	495.59	443.60	362.39	346.89	0.00	398.13	406.16	363.44
8	498.56	464.59	275.92	337.29	350.04	379.61	398.13	0.00	486.60	402.21
9	464.97	569.54	582.96	538.33	436.80	389.67	406.16	486.60	0.00	449.59
10	441.18	521.14	487.70	461.32	326.17	349.61	363.44	402.21	449.59	0.00

Table 3: Matriz de similaridade para as primeiras 10 obras (vetores com dimensionalidade 100)

	1	2	3	4	5	6	7	8	9	10
1	0.00	574.38	615.91	562.38	449.66	371.01	421.97	494.98	458.30	438.20
2	574.38	0.00	501.49	507.95	499.24	497.75	518.09	475.00	582.71	533.07
3	615.91	501.49	0.00	317.41	431.67	499.03	515.04	281.45	597.52	503.37
4	562.38	507.95	317.41	0.00	410.47	467.27	447.66	334.01	541.67	464.62
5	449.66	499.24	431.67	410.47	0.00	353.55	367.06	350.18	437.58	334.57
6	371.01	497.75	499.03	467.27	353.55	0.00	342.17	385.23	388.76	348.52
7	421.97	518.09	515.04	447.66	367.06	342.17	0.00	407.51	404.75	361.93
8	494.98	475.00	281.45	334.01	350.18	385.23	407.51	0.00	494.31	408.39
9	458.30	582.71	597.52	541.67	437.58	388.76	404.75	494.31	0.00	451.02
10	438.20	533.07	503.37	464.62	334.57	348.52	361.93	408.39	451.02	0.00

Table 4: Matriz de similaridade para as primeiras 10 obras (vetores com dimensionalidade 200)