Resumo Artigo - "Learning Word Vectors for Sentiment Analysis"

Patricia Sayuri (NUSP: 11338327) e Luis Alvaro Correia (NUSP: 745724)

June 27, 2019

1 Introdução

A representação de palavras tem um papel crítico em vários sistemas de processamento de linguagem natural (*NLP processing Systems*). A melhor forma de representação é a vetorial, onde é possível encapsular diferenças e similaridades através de distâncias e medidas angulares desses vetores em um espaço multi-dimensional.

Neste artigo, será apresentado um modelo que identifica as similaridades semânticas e de sentimento entre palavras através de um modelo probabilístico não-supervisionado sobre documentos. Isso será feito através da representação vetorial das palavras que possibilitará capturar essas semelhanças. Em nosso experimento será possível alavancar esses sentimentos através das anotações que consumidores efetuam para filmes, produtos etc.

Após a apresentação do modelo, forneceremos algums exemplos ilustrativos dos vetores contruídos bem como eles são aplicados para executar tarefas de classificação em diversos níveis. Para isso será utilizado primordialmente as revisões contidas na base de dados do IMDB (*Internet Movie Database*).

2 Trabalhos Relacionados

O modelo apresentado neste artigo compartilha os fundamentos do modelo LDA (*Latent Dirichlet Allocation* (Blei et al.,(2003)) [1] que assume que um documento é composto de uma mistura de tópicos onde o adaptamos para que ele seja fatorado em forma de vetores e cujas representações permitam identificar palavras ao invés de tópicos.

Para isso é importante entender a organização e definições que serão utilizadas a partir deste ponto.

Formalmente definimos os seguintes termos:

- Uma *palavra* é a unidade básica de dados, definido como um item num vocabulário e é indexada por {1,...,V}. Ela é representada por vetores que tem um único componentes igual a "1" e todos os outros iguais a "0". Podemos formalmente definir então que a v-ésima palavra de um vocabulário é representada por um vetor de dimensão V tal que $w^v = 1$ e $w^u = 0$ para $u \neq v$;
- Um *documento* é uma sequência de N palavras denotadas por $\mathbf{w} = (w_1, w_2, ..., w_N)$ onde w_n é a n-ésima palavra na sequência;
- Uma *coleção* é um conjunto de M documentos denotados por $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$.

Através de modelos vetoriais (VSM) é possível implementar o aprendizado semântico das palavras, porém, de modo a inserir também informações referentes ao sentimento é necessário incorporar pesos no modelo vetorial através de métodos supervisionados.

3 O Nosso Modelo

Com o objetivo de capturar similaridades semânticas entre as palavras, derivaremos um modelo probabilístico de documentos que aprende a partir da representação vetorial de palavras. Neste contexto, a componente *sentimento* de nosso modelo utilizará os comentários para restringir palavras que expressem os mesmo sentimentos e assim estimar os parâmetros do modelo.

3.1 Captura de Similaridades Semânticas

O modelo probabilístico construído considera a probabilidade de um documento d condicionado à um parâmetro multidimensional de *mistura* aleatória θ de tópicos, onde cada tópico é caracterizado por uma distribuição de palavras. Neste modelo assumimos que cada palavra $w_i \in d$ é condicionalmente independente do parâmetro θ.

Desda forma, a probabilidade de um documento é:

$$p(d) = \int p(d, \theta) d\theta = \int p(\theta) \prod_{i=1}^{N} p(w_i|\theta) d\theta \tag{1}$$

onde N é o número de palavras em d e w_i é a i-ésima palavra em d. Este modelo é semelhante ao apresentado em [1].

Aqui definimos $p(w_i|\theta)$ como um modelo log-linear com parâmetros R e b onde R é a matriz de representação onde cada palavra w é representada por um vetor de dimensão β (*one-on vector*) no vocabulário V e pode ser expressa por $\phi_w = Rw$ que terá "1" apenas na coluna correspondente a esta palavra na matriz R.

De maneira análoga, o parâmetro $\theta \in \mathbb{R}^\beta$ é um vetor β-dimensional que pondera cada uma das β-dimensões das representações das palavras.

Nesta abordagem, modelamos a probabilidade das palavras condicionado à uma mistura de tópicos θ. Podemos entender as entradas do vetor de palavra ϕ como sendo o grau de associação desta palavra com respeito a cada um dos tópicos e o parâmetro θ aleatório definindo os pesos desses dos tópicos.

O próximo passo é derivar a função de máxima verossimilhança deste modelo, dado um conjunto de documentos D.

Aqui assumimos que os documentos $d_k \in D$ são i.i.d., portanto, temos:

$$\max_{R,b} p(D; R, b) = \prod_{d_k \in D} \int p(\theta) \prod_{i=1}^{N_k} p(w_i|\theta; R, b) d\theta \tag{2}$$

por aproximação, a equação 2 pode ser escrita como:

$$\max_{R,b} \prod_{d_k \in D} p(\hat{\theta}_k) \prod_{i=1}^{N_k} p(w_i|\hat{\theta}_k; R, b) \tag{3}$$

Após simplificações e normalização da matriz R, obtemos a equação de aprendizado, cujos hyper-parâmetros λ, β e v maximizam a função em relação a R e b, e que pode ser ser escrita como:

$$v\|R\|_F^2 + \sum_{d_k \in D} \lambda \left\| \hat{\theta}_k \right\|_2^2 + \prod_{i=1}^{N_k} log[p(w_i|\hat{\theta}_k; R, b)] \tag{4}$$

3.2 Captura de Sentimento das Palavras

O modelo apresentado na seção anterior produz representações semelhantes para palavras que ocorrem juntas em determinados documentos. É necessário, então, definir como capturar os sentimentos que estas palavras expressam.

Para isso, definiremos uma categoria s de sentimentos ao qual a representação vetorial da palavra ϕ_w deverá predizer, através de uma função preditora, o sentimento a ela associado ou seja:

$$\hat{s} = f(\phi_w) \tag{5}$$

Em associação com rating fornecido na revisão do usuário, que pode ser estabelecida como uma escala tal que $s \in [0, 1]$ utilizaremos como função preditora a regressão logística, com:

$$p(s = 1|w; R, \psi) = \sigma(\psi^T \phi_w + b_c) \tag{6}$$

onde $\sigma(x)$ é a função logística, $\psi \in \mathbb{R}^{\beta}$ são os pesos da função logística e b_c um viés introduzido pelo classificador.

Desta forma, a função de classificação obtida define um hyperplano sobre o qual calcularemos a probabilidade do espaço vetorial das palavras do documento em relação a este hyperplano.

O aprendizado da função de classificação da equação 6 será então feita através de um conjunto de documentos D cujos labels de sentimentos s_k são conhecidos, ou seja, devemos maximizar a função:

$$\max_{R,\psi,b_c} \sum_{k=1}^{|D|} \sum_{i=1}^{N_k} log[p(s_k|w_i; R, \psi, b_c)] \tag{7}$$

3.3 O Aprendizado

O objetivo de aprendizado pode ser expressado pela soma das quantidades apresentadas nas seções 3.1 e 3.2 como segue:

$$v\|R\|_F^2 + \sum_{d_k \in D} \lambda \left\| \hat{\theta}_k \right\|_2^2 + \prod_{i=1}^{N_k} log[p(w_i|\hat{\theta}_k; R, b)] + \sum_{k=1}^{|D|} \frac{1}{|S_k|} \sum_{i=1}^{N_k} log[p(s_k|w_i; R, \psi, b_c)] \tag{8}$$

onde $|S_k|$ representa o número de documentos no dataset com o mesmo s_k (i.e. $s_k < 0.5$ e $s_k \geq 0.5$) e introduzimos a quantidade $\frac{1}{|S_k|}$ para reduzir o desbalanceamento dos ratings das avaliações eventualmente presentes em cada documento do dataset.

Desta forma, a solução para nosso problema consistirá na maximização da equação 8 em relação a R, b, ψ and b_c e se reduz a um problema de otimização não-convexa, portanto utilizaremos o método *maximum a posteriori* (MAP) para um $\hat{\theta}$ fixo.

4 Experimentos

Avaliaremos nosso modelo em relação a classificação de documentos e sentenças sobre o domínio das revisões de filmes. Para isso utilizaremos os dados disponíveis de aproximadamente 25.000 revisões no IMDB, considerando até 30 reviews por filme.

Algumas considerações importantes:

- Construímos um vocabulário de 5.000 palavras mais frequentemente encontradas;
- Não fizemos remoção de *stop words* devido a algumas delas denotarem sentimentos;

- Redução à raiz semântica não foi utilizado pois o modelo identifica como similares palavras com mesma raiz. Em contrapartida, foram inseridas expressões como "!" e ":-)" pois são indicativas de sentimentos;
- Os ratings do IMDB foram linearmente mapeados no intervalo $[0, 1]$;
- Para a componente semântica do modelo foram utilizadas as revisões sem título;
- A avaliação qualitativa do modelo considera a similaridade entre duas palavras através do co-seno entre suas representações vetoriais, onde palavras com a mesma orientação têm co-seno de similaridade 1; duas palavras ortogonais entre si possuem similaridade *zero*, e duas palavras diametralmente opostas têm similaridade -1, ou seja $S(\phi_w, \phi_{w'}) = \frac{\phi_w^T \phi_{w'}}{\|\phi_w\| \|\phi_{w'}\|}$.

Em comparação com outros métodos, a inclusão de sentimentos mostra que o modelo evita algumas confusões pois aproxima as palavras com a mesma polaridade de sentimentos tornando, mesmo o aprendizado não-supervisionado, bastante promissor.

Nosso modelo foi testado contra alguns algoritmos VSM, tais como LDA (*Latent Dirichlet Allocation*)[1], Variantes Ponderadas (*Weighting Variants*)[2], Classificação de Polaridade (*Polarity Classification*)[3] e LSA (*Latent Semantic Analysis*)[4].

O objetivo principal do modelo é identificar a *polaridade* das avaliações que consiste em decidir se uma revisão pode ser classificada como *positiva* ou *negativa*.

Nosso modelo apresentou performance superior em relação aos algoritmos comparados, sendo que em relação à acuracia, a performance foi semelhante.

Além disso, o modelo também foi testado para identificar *subjetividade* das avaliações, ou seja, uma tarefa substancialmente diferente da tarefa principal e que consiste em classificar se uma dada sentença é *subjetiva*, ou seja, baseada apenas na opinião do avaliador; ou se é *objetiva*, quando ela é baseada em fatos. De forma semelhante, o modelo apresentado neste documento teve uma performance superior em relação aos outros VSM's.

5 Conclusão

O modelo que incorpora também sentimentos além de características semânticas de palavras mostra-se promissor em comparação aos VSM's propostos. Alem de um fundamento probabilístico, nosso modelo é parametrizado como um modelo log bi-linear, em consonância com estudos de sucesso em modelos linguísticos.

Além disso, outra inovação deste modelo é a representação de palavras que incorpora sua representação vetorial ao invés de capturar seus tópicos latentes o que proporciona uma performance superior aos outros modelos.

Ao estender o aprendizado não-supervisionado deste modelo, foi possível incorporar informações de sentimento relevantes, fartamente disponíveis em diversos meios, como redes sociais, sites de pesquisa e que contribuem para uma melhor performance e flexibilidade para utilização em outras áreas de pesquisa.

References

[1] Michael I. Jordan David M. Blei, Andrew Y. Ng. Latent dirichlet allocation. *Journal of Machine Learninig Research 3*, 2003.

[2] G. Paltoglou and M. Thelwall. A study of information retrieval weighting schemes for sentiment analysis. *In Proceedings of ACL*, 2010.

[3] B. Pang and L.Lee. A sentimental educstion: sentiment analysis using subjectivity summarization based on minimum cuts. *In Proceedings of ACL*, pages 271–278, 2004.

[4] G. W. Furnas T. K. Landauer S. Deerwester, S. T. Dumais and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391—407, September 1990.