Modelos estatísticos para classificação de mensagens em redes sociais: uma análise das menções à Unicamp no Twitter

Palavras-Chave: Twitter, Biclusterização, Análise de Agrupamentos

Autores:

Antonio Vítor Ribeiro [IMECC]

Prof. Dr. Guilherme Vieira Nunes Ludwig (orientador) [IMECC]

INTRODUÇÃO:

Redes sociais são um dos meios de comunicação modernos mais utilizados entre a população mundial. O uso de redes sociais se estende inclusive para interação e comunicação política (Zhang et al., 2017, entre outros). De uma maneira geral, as opiniões compartilhadas em redes sociais tipicamente são orientadas por poucos agentes representativos que tendem a influenciar seu círculo de seguidores (Webster, 2014; Zhang et al., 2017).

Há diversos aspectos de análise de dados em redes sociais que precisam ser considerados, como o processamento de texto (veja, e.g., Silva e Ribeiro, 2009) e critérios adequados de classificação (Schölkopf e Smola, 2002) e agrupamento (Bouveyron et al., 2019). Por outro lado, a estrutura da rede social também pode ser estudada, identificando seus atores principais, agrupamentos e cliques (Wasserman e Faust, 1994).

Para a realização de análises com aspectos simultâneos, como objetos e atributos, analises de *clusterização* simples não são o suficiente, e para isso, existe o conceito de *biclusterização*. Desse modo, a *biclusterização* (Hartigan, 1972) é uma técnica de agrupamento para uma base de dados, em que as linhas (ou indivíduos) são agrupados simultaneamente com as colunas (ou atributos). Assim, linhas agrupadas exibem características similares entre grupos de atributos, e vice-versa.

METODOLOGIA:

Coleta dos dados

banco de dados de mais de 120 mil tweets, e análises preliminares dos mesmos. As análises foram realizadas em linguagem R, e o conteúdo foi acessado através do API do Twitter, usando o pacote twitteR (Gentry, 2015; R Core Team, 2020). Após os dados serem recolhidos, eles são processados e os atributos são extraídos. Entre os atributos, temos o usuário que fez o tweet, a data em que o tweet foi coletado, hyperlinks e contagem de interações e retweets.

Anteriormente, na extensão do projeto, os tweets foram reunidos em todas as semanas, para que pudéssemos realizar o agrupamento simultâneo de usuários e termos (correspondendo neste caso a linhas e colunas, no contexto de *biclusterização*). Um dicionário com os termos mais frequentes foi criado, entretanto, nem todos os termos mais frequentes foram utilizados para as análises, pois não apresentavam valor algum, por serem palavras de dias ou situações especificas, como por exemplo termos da “Semana da Copa da Engenharia da Unicamp”. Essa inspeção foi realizada ad hoc, através da análise uma a uma das palavras mais comuns.

Entretanto, devido a enorme quantidade de dados coletados, apenas o pacote ggplot2 (H. Wickham, 2016) suportou realizar a análise de todos os tweets juntos, em relação ao dicionário criado. Desse modo, todos os retweets foram retirados da análise, fazendo com que o número total caísse quase pela metade, e possibilitando uma análise que favorece contas que normalmente não têm grande visibilidade, como de portais de notícias ou influencers.

O algoritmo de biclustering

Biclusterização pode ser denominada por técnicas que buscam agrupar e concomitantemente a uma base de dados organizada em indivíduos e atributos. Na base de dados exibida na Figura 1, temos que as linhas podem ser agrupadas usando um procedimento hierárquico, baseado na distância entre atributos, neste caso, o agrupamento é considerado global. Para o procedimento de *biclustering*, um cluster de atributos e linhas similares indicado em tom cinza escuro representa o agrupamento com natureza local.

1	1	2	2	2	1	5	5	4	4	2
1	1	2	1	2	1	5	5	4	5	1
2	2	2	1	1	1	3	4	5	5	1
5	4	5	4	4	5	2	2	3	1	1
4	5	5	4	3	3	1	1	2	4	1
4	4	5	4	3	1	5	4	4	5	1
5	5	4	5	1	2	3	1	3	4	1

Figura 1: Base de dados para exemplo

Neste trabalho utilizaremos o *bicluster* que reorganiza linhas e colunas na matriz encontrando linhas e colunas com valores semelhantes.

O seguinte algoritmo apresenta uma construção de *biclusters* através da organização entre linhas e colunas. Tome X uma matriz $n \times p$ em que tenho n mensagens (ou tweets) codificadas em um dicionário de p termos mais frequentes. A ocorrência ou não de mensagens é codificada em $0s$ e $1s$.

```
Tweet #1:

Display Name: bárbara!
Username: @studybarbara
Date: Apr 19
Text: a unicamp é linda, fico encantada toda vez que meu namorado me manda foto das aulas que
ele tem

Tweet #2:

Display Name: zépa 13
Username: @zepamoraes
Date: Apr 19
Text: obrigado pela sobremesa maravilhosa de hoje bandejão unicamp

Tweet #3:

Display Name: Ricardo Miranda Martins
Username: @rmiranda00
Date: Apr 19
Text: Esses textos do Martinez no Jornal da Unicamp estão excelentes.

Tweet #4:

Display Name: Ivan
Username: @ivanknobel
Date: Apr 18
Text: feijoada e strogonoff no bandeco em um espaço de 3 dias tem que ser uma cortina de
fumaça pra alguma coisa muito ruim acontecendo na unicamp
```

A menos do termo **unicamp**, comum a todas as mensagens, e de palavras de ligação / *stopwords* e expressões comuns, podemos extrair atributos (palavras lematizadas), por exemplo $A =$ [linda/maravilhosa/excelente], $B =$ [bandejão/bandeco], $C =$ [aula], e produzir a matriz de ocorrências X:

	A	B	C
1	1	0	1
2	1	1	0
3	1	1	0
4	0	1	0

A organização dos *biclusteres* é determinada pela maximização da seguinte função objetivo:

$$s(X, r, c) = b(r, c) \sum_{ijk} x_{ij} r_{ik} c_{jk} \,,$$

em que $r_{ik} \in \{0, 1\}$ é uma etiqueta de pertencimento da linha i (usuário) ao cluster k, e $c_{jk} \in \{0, 1\}$ é uma etiqueta da coluna j (atributo, ou termo) ao cluster k. O termo b é uma função que penaliza a diferença de tamanho entre clusteres: isto é, $b(r, c) = \alpha(max\, S_k - min\, S_k)$ onde

$$S_k = \left(\sum_i r_{ik}\right) \cdot \left(\sum_j c_{jk}\right).$$

O termo α é uma constante de regularização. A busca de etiquetas pode ser realizada através de um procedimento de *simulated annealing*, onde cada par de linhas / colunas tem propostas de etiquetas trocadas aleatoriamente; a troca é aceita com probabilidade 1 se aumenta s, e aceita com alguma probabilidade decrescente (em função de α) se diminui s. No exemplo anterior, a configuração de *biclustering* encontrada é

	C	A	B
1	1	1	0
2	0	1	0
3	0	1	1
4	0	0	1

Dentre os diferentes métodos para a *biclusterização*, está técnica apresenta uma avaliação mais simples, pois procura por agrupamentos em que seus objetos tenham um comportamento semelhante. Por mais que seja um método com avaliação mais simples, ele requer um pré-processamento dos dados, para que os mesmos não apresentem ruídos para que melhores agrupamentos sejam formados, assim como realizado em Symeonidis et al. (2007), que também transformou seus dados em uma matriz binária.

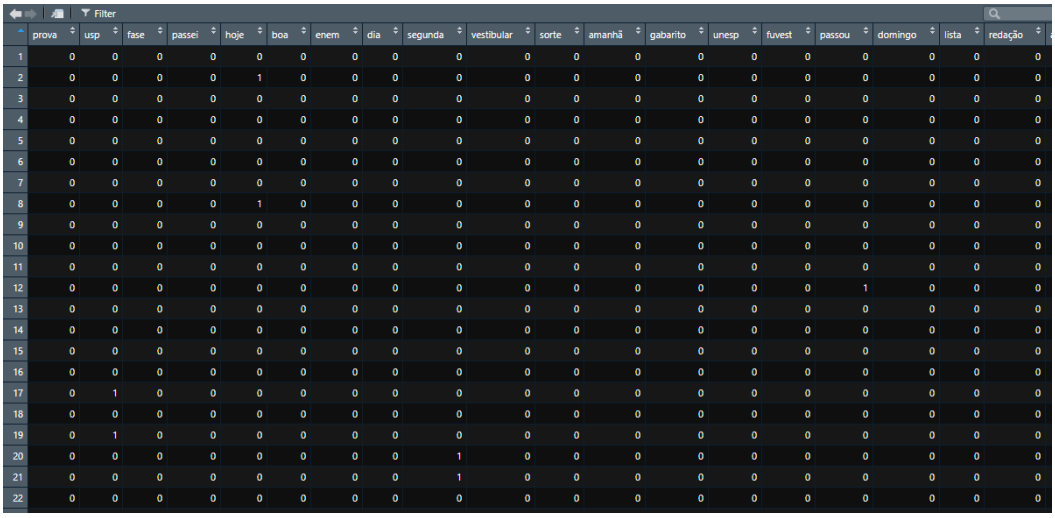
RESULTADOS E DISCUSSÃO:

A Figura 2 apresenta o dicionário de termos mais frequentes entre as semanas analisadas, desconsiderando todos os retweets. É possível notar que termos que remetem a faça de participar de vestibulares são os termos mais frequentes quando os usuários tweetam mencionando a palavra Unicamp. Um motivo para isso é que, ao retirar os retweets, muitos termos não são mencionados no cotidiano dessa população, que tweetam sobre a Unicamp e, também, por se tratar de uma rede social, muitos termos e palavras são escritos de forma abreviadas, requerendo assim maiores esforços no processamento, tornando assim a coleta desses termos menos eficaz.

```
   prova      usp     fase   passei     hoje      boa    enem      dia  segunda vestibular
    3690     1605     1482      977      899      817     786      758      701       634
   sorte  amanhã  gabarito   unesp   fuvest   passou  domingo   lista  redação      anos
     533      517      511      462      457      428     369      322      226       215
economia juliette  honoris instituto  semana   jarbas   conta     tava      mês professor
     204      149       96       85       78       78      76       63       55        42
  passar
      35
```

Figura 2: Dicionário de termos mais frequentes sem retweets

A Figura 3 apresenta a matriz de tweets e termos frequentes, sendo as linhas cada tweet coletado e as colunas cada termo do dicionário criado. O número sim significa a presença do respectivo termo no texto do tweet em questão, analogamente o 0 representa a ausência. A matriz conta com quase 55 mil linhas e 31 colunas. A grande quantidade de 0 na matriz é razoável, uma vez que a população de alguma forma relacionada com a Unicamp é maior dos vestibulandos, dessa forma, como as palavras mais frequentes foram relacionadas a eles, a grande parte dos tweets não terá tais termos no texto do tweet.



	prova	usp	fase	passei	hoje	boa	enem	dia	segunda	vestibular	sorte	amanhã	gabarito	unesp	fuvest	passou	domingo	lista	redação
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 3: Matriz tweets X termos mais frequentes sem retweets

A Figura 4 apresenta um *heatmap*, criado com o pacote pheatmap (Kolde, 2019), em relação a Figura 3. Devido a dimensão dos dados, uma amostra aleatória de 250 tweets foi analisada. O gráfico apresenta analises que valem a pena serem observadas, como a criação de 3 possíveis agrupamentos entre os tweets, representados pela organização do dendrograma vertical. Sobre esses possíveis agrupamentos, temos que na parte inferior do gráfico, os tweets apresentam nenhum dos termos mais frequentes em seus textos, podendo ser entendido como funcionários, população externa a Unicamp ou até mesmo tweets específicos de alunos, já na parte superior, temos a presença de dois possíveis agrupamentos. De acordo com os termos e a *clusterização* entre eles, é possível entender o primeiro agrupamento como sendo de vestibulandos, *tweetando* sobre a prova e vestibulares, bem como as coisas que se relacionam com isso, como o termo “sorte”, remetendo a “boa sorte”, enquanto que o agrupamento abaixo pode representar alunos da universidade, mencionando provas, professores e assuntos internos dos campi.

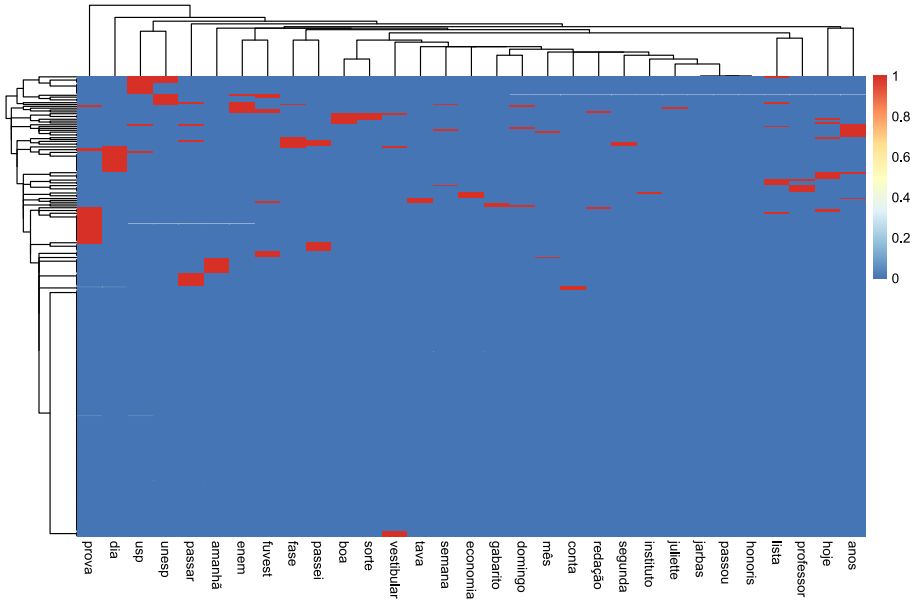


Figura 4: Heatmap de uma amostra aleatória de 250 tweets da matriz tweets X termos mais frequentes sem retweets

CONCLUSÕES:

Levando em consideração esses aspectos, uma análise sem os retweets acaba apresentando mais relevantes, pois os termos frequentes estavam sendo apenas de notícias ou tweets virais muito específicos da universidade, que não tinham valores. Assim, a média de tweets semanais esteve próxima a 1,5 mil tweets.

O dicionário e o *heatmap* exibem que, muito provavelmente, os usuários que mais mencionam a Unicamp no Twitter são vestibulandos, devido a configuração dos *clusters* e aos termos mais frequentes. O *heatmap* ainda apresenta 3 possíveis agrupamentos, sendo esses de vestibulandos, alunos que *tweetam* sobre coisas mais recorrentes da universidade e alunos e funcionários que *tweetam* coisas especificas, não tendo muitos termos comuns.

BIBLIOGRAFIA

BOUVEYRON, C., CLEUX, G., Brendan Murphy, T. e Raftery, A. E. (2019). ***Model-based clustering and classification for data science***. Cambridge University Pr.

HARTIGAN, John A. Direct clustering of a data matrix. **Journal of the american statistical association**, v. 67, n. 337, p. 123-129, 1972.

KEARNEY, M. W. (2019). rtweet: Collecting and analyzing Twitter data, Journal of Open Source Software, 4, 42. 1829. doi:10.21105/joss.01829 (R package version 0.7.0)

Kolde R (2019). _pheatmap: Pretty Heatmaps_. R package version 1.0.12, <https://CRAN.R-project.org/package=pheatmap>.

SCHÖLKOPF, B. e SMOLA, A. J. (2002). ***Learning with kernels: support vector machines, regularization, optimization, and beyond***. MIT press, Cambridge, Mass.

SILVA, C. e RIBEIRO, B. (2009). ***Inductive inference for large scale text classification: kernel approaches and techniques***. Springer, New York.

SYMEONIDIS, Panagiotis et al. Nearest-biclusters collaborative filtering with constant values. In: **Advances in Web Mining and Web Usage Analysis: 8th International Workshop on Knowledge Discovery on the Web, WebKDD 2006 Philadelphia, USA, August 20, 2006 Revised Papers 8**. Springer Berlin Heidelberg, 2007. p. 36-55.

WASSERMAN, S. e FAUST, K. (1994). ***Social network analysis: methods and applications***. Cambridge University Pr.

WEBSTER, J. G. (2014). ***The marketplace of attention: how audiences take shape in a digital age***. MIT Press, Cambridge, Mass.

WICKHAM et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

ZHANG, Y., WELLS, C., WANG, S. e ROHE, K. (2017) **Attention and amplification in the hybrid media system: The composition and activity of Donald Trump's Twitter following during the 2016 presidential election**. *New Media & Society, v. 20, no. 9, pp. 3161-3182.*