IPEA - DISET

Indicador de Tecnologias Emergentes

Fred Guth*
27 de agosto de 2019

1 OBJETIVO

Antecipar tecnologias emergentes com maior potencial de impacto é essencial para um bom planejamento de políticas públicas. Usualmente, essa atividade é bastante qualitativa, se baseando no conhecimento de especialistas e na análise de patentes e artigos científicos. Entretanto, redes sociais passaram a expressar, em tempo real, a opinião de diferentes agentes econômicos, incluindo empresas, institutos de pesquisa e publicações de especializadas em tecnologias emergentes; tornando possível mensurar quantitativamente, pelo número de menções, o julgamento agregado dessa multidão de agentes do potencial de tecnologias (assumindo-se que os agentes mencionam mais as tecnologias em que mais vêem potencial). É um fenômeno bem relatado (Surowiecki, 2004) que julgamentos agregados deste tipo produzem boas previsões.

Neste projeto, usamos processamento de linguagem natural para criar um indicador quantitativo para tecnologias emergentes mencionadas no Twitter por contas previamente selecionadas.

2 METODOLOGIA

O processo de construção do **Indicador de Tecnologias Emergentes (ITE)** abrange as seguintes etapas:

1. Captura dos tweets[1]
2. Análise exploratória dos dados
3. Limpeza dos dados e definição dos tópicos
4. Construção da matriz de frequências Semestre-Tópico
5. Detecção estatística de anomalias
6. Geração do Indicador

*consultor contratado.

[1] A curadoria das contas e a captura dos *tweets* foram realizadas previamente e fogem do escopo do presente projeto.

2.1 ANÁLISE EXPLORATÓRIA DOS DADOS

A análise exploratória de dados (AED) é uma abordagem de análise com objetivo de capturar um panorama geral dos dados com métodos visuais. No presente projeto, a AED não foi estensiva, até mesmo porque a maior variabilidade dos dados está no texto dos tweets o que não é capturado pela mesma. Entretanto, a AED foi importante para identificar potenciais problemas com diferentes abordagens estatísticas. Os resultados da AED serão apresentados na seção 3.

2.2 LIMPEZA DOS DADOS E DEFINIÇÃO DOS TÓPICOS

2.2.1 LIMPEZA

A limpeza dos dados tem como objetivo padronização e a remoção de termos anômalos indesejados na análise. Todas as ações de limpeza são opcionais e podem ser *desligadas* a partir de um arquivo de configuração (ver **??**).

- remoção de acentos
- remoção de *hashtags* e menções a contas do twitter: apesar de hashtags serem um sinal forte de importância no Twitter, achamos importante dar a opção de fazer a análise sem hashtags.
- remoção de URLs
- remoção de números: apenas palavras contendo apenas números, por exemplo "2017"é removido, enquanto "23andme"não é.

2.2.2 DEFINIÇÃO DOS TÓPICOS

Qualquer palavra usada em um *tweet* do corpus é, inicialmente, considerada um tópico, uma vez que, a priori, não temos como saber que uma palavra se trata de tecnologia. Pior do que isso, como analisamos também bi-gramas, combinações de duas palavras, no limite o número de tópicos pode ser quadrático em relação ao tamanho do vocabulário do corpus.

Tendo em vista que nosso corpus contém apenas mensagens curtas, o contexto em que termos tecnológicos aparecem não são muito identificáveis. Em outras palavras, queremos encontrar um pequeno sinal em um grande mar de ruído. Essa característica do problema nos leva a focar primeiro em identificar anomalias em geral, que podem ou não ser tecnológicas, deixando o problema da categorização dos tópicos como secundário.

Um primeiro passo é, então, reduzir o tamanho do vocabulário considerado. Parte disso é feito pela própria limpeza, mas consideramos também que palavras devem ter um número mínimo de menções-dia[2]. Além desses filtros, também estabelecemos um limite (*dict_size*) no tamanho do dicionário, pegando apenas as *dict_size* palavras com mais menções-dia no corpus.

Também filtramos *stop words*, que são palavras comuns. Todas as principais bibliotecas de NLP possuem uma lista de *stop words* do idioma inglês. Utilizamos a lista da biblioteca NLTK (Loper and Bird, 2002) e adicionamos uma lista própria de *stop words* que pode ser alterada no arquivo de configuração. Tal lista foi criada especificamente para o desafio deste projeto e parte da constatação que mesmo contas especializadas engajam-se em assuntos da cultura e capturam o zeitgeist ao longo do tempo.

[2] menção-dia é frequencia de menções em diferentes dias, ou seja, várias menções no mesmo dia contam como uma.

2.3 Matriz de frequências Semestre-Tópico

2.4 Detecção estatística de anomalias

2.5 Geração do Indicador

3 Dados

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt.

Figura 3.1: Tweets por fonte.

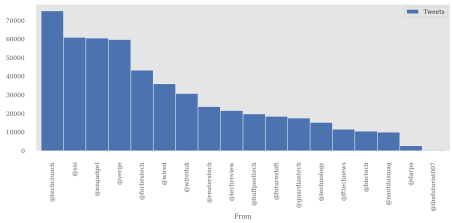


Figura 3.2: Tweets por dia.

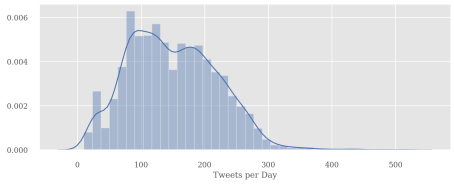
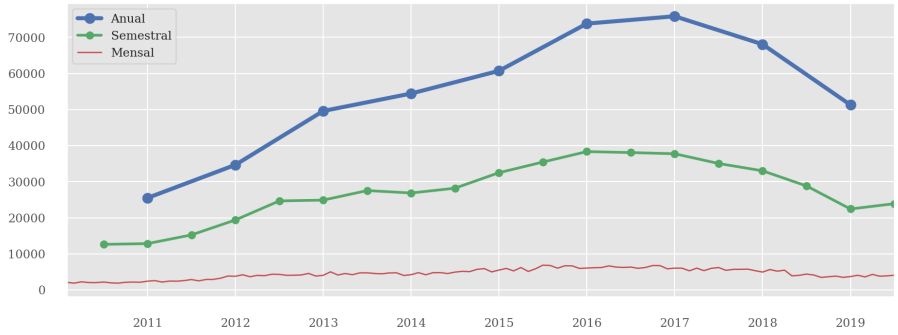


Figura 3.3: Tweets por período.



4 Código

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus:

Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies

4.1 HEADING ON LEVEL 2 (SUBSECTION)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit.

(4.1)

Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem.

4.1.1 HEADING ON LEVEL 3 (SUBSUBSECTION)

Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim.

HEADING ON LEVEL 4 (PARAGRAPH) Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim.

5 PRODUTO

5.1 EXAMPLE FOR LIST (3*ITEMIZE)

- First item in a list
  - First item in a list
    - First item in a list
    - Second item in a list
  - Second item in a list
- Second item in a list

5.2 EXAMPLE FOR LIST (ENUMERATE)

1. First item in a list
2. Second item in a list
3. Third item in a list

REFERÊNCIAS

Loper, E. and S. Bird (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations 296*.