Técnicas de Aprendizado de Máquina Aplicadas a Negócios

**Laboratório Análise de Textos**

**Introdução**

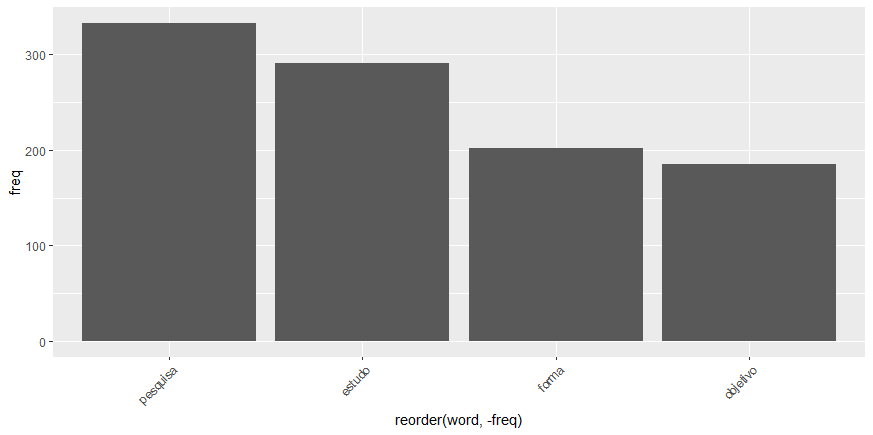
O presente relatório foi realizado tendo como base o script fornecido pelo Professor, como foi feito em laboratórios anteriores. No entanto, não conseguiu-se integrar à base de textos os arquivos em PDF do SemeAD 2019, mesmo após a consulta de fóruns na internet (como Stack Overflow) e diversas tentativas. Assim, utilizamos apenas os arquivos que estavam em formato “.txt”, que no caso eram aqueles referentes aos textos do SemeAD 2017, como foi feito no script apresentado no laboratório também.

Não obstante, acreditamos que o objetivo principal do laboratório tenha sido alcançado (i.e., aplicar a técnica de aprendizado em questão e interpretar os resultados produzidos).

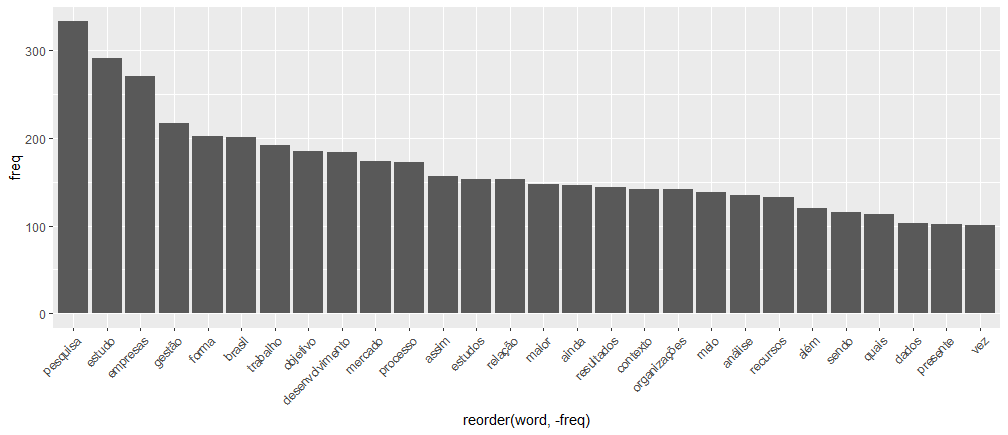
**Frequência de Termos**

Quanto aos termos mais frequentes nos documentos analisados, pode-se se deparar com um dilema quanto à definição da variável “sparse” da função “removeSparseTerms”.

Utilizando-se do valor 0.3, obtém-se a o seguinte resultado abaixo:



Assim, apenas essas 4 palavras destacadas no gráfico acima se encaixam nos critérios de o quão esparsas elas estão (o 0.3) e da frequência mínima, no caso de 100. Tendo em vista que tal resultado não permite a extração de informações tão interessantes sobre o corpus, definiu-se a variável “sparse” para 0.6, obtendo-se o seguinte gráfico:



Assim, pode-se identificar palavras interessantes para a análise do conjunto de textos, como “Brasil” e “processo”. Nota-se, no entanto, que seria necessária uma limpeza mais rigorosa das palavras, porque termos pouco significativos, como “assim”, “vez” e “sendo”, estão sendo considerados também.

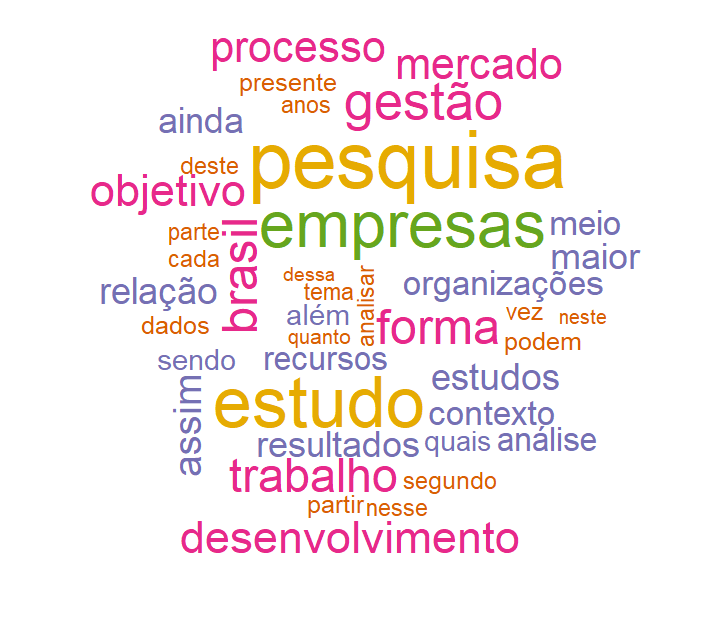
**Correlação de Palavras**

Tendo-se em vista as duas palavras destacadas anteriormente, “Brasil” e “processo”, utilizou-se a função “findAssics” para encontrar quais palavras são comumente utilizadas juntamente com elas.

Quanto à “Brasil”, as palavras “país” e “praticados” são as que têm mais correlação e parecem fazer sentido. Tendo a palavra “processo” em mente, descobriu-se que ela está associada às palavras “condição” e “padronização”, principalmente.

**Nuvem de Palavras**

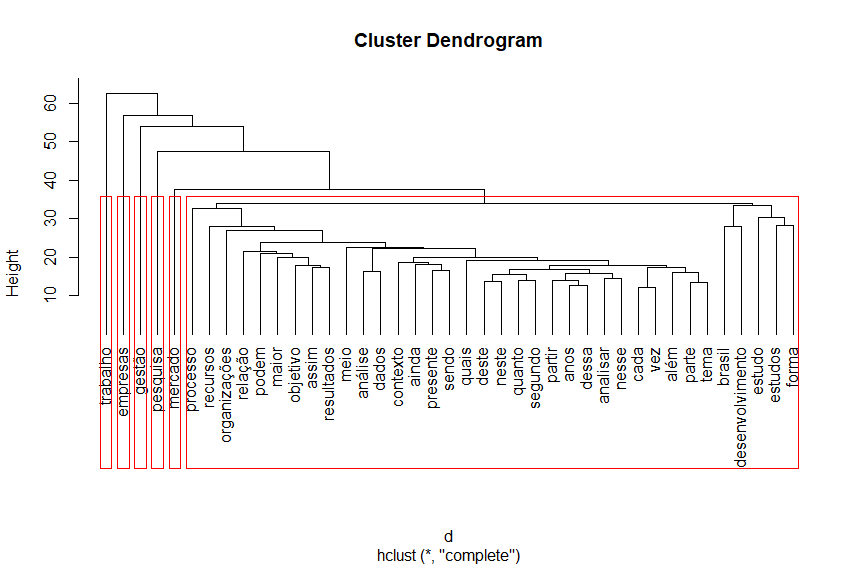
Obteve-se ainda a seguinte nuvem de palavras colorida considerando aquelas palavras que ocorreram pelo menos 50 vezes nos textos analisados no presente laboratório:



Como pode ser visto na imagem acima, a nuvem mostra as palavras mais frequentemente encontradas nos textos analisados, sendo que quanto maior o tamanho da palavra, maior sua frequência. Sendo assim, palavras como “pesquisa”, “empresas” e “estudo” apareceram nos textos analisados com maior frequência e pelo menos 50 vezes.

**Análise de Agrupamento**

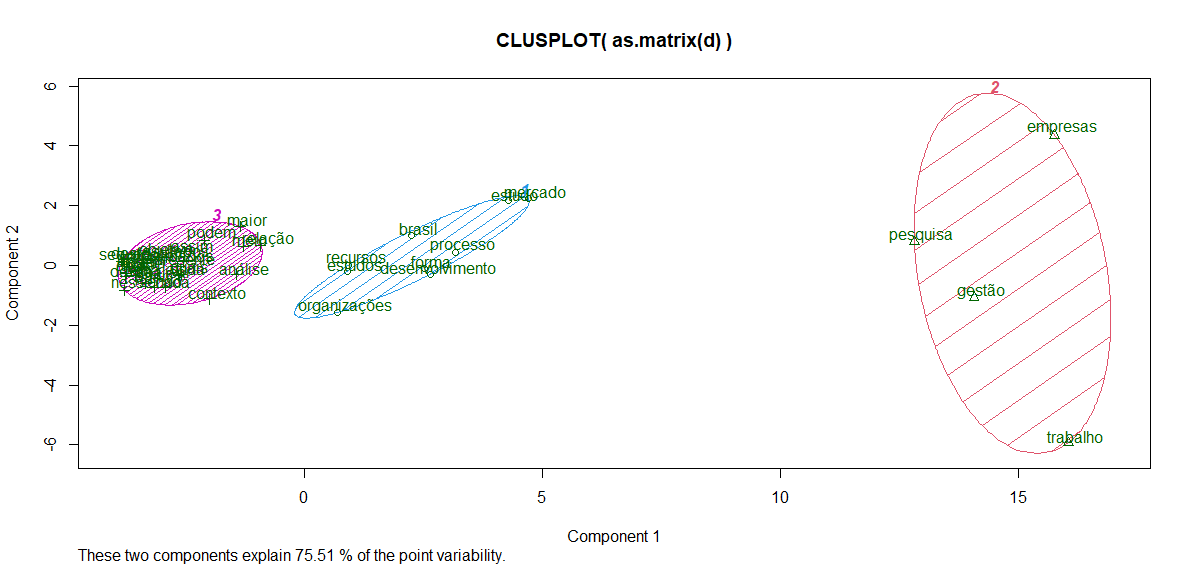
Por fim, o dendrograma abaixo representa o agrupamento hierárquico das palavras a partir da distância euclidiana entre elas, tendo como base os textos analisados pelo algoritmo e considerando a formação de 6 grupos.



A partir do dendrograma, é possível observar que as palavras “trabalho”, “empresas”, “gestão”, “pesquisa” e “mercado” foram agrupadas individualmente, já as demais palavras formaram um cluster com uma maior quantidade de valores, incluindo diversas palavras que não contém tanto significado por si só, como “partir”, “neste”, “além” e “vez”. É interessante pontuar isso tendo em vista que, no tópico seguinte, poderá ser observado que a maioria das palavras que foram agrupadas separadamente pelo algoritmo de agrupamento hierárquico serão agrupadas conjuntamente quando utiliza-se a técnica de agrupamento por partição (k-means).

**Agrupamento por Partição**

Utilizando-se a técnica de agrupamento por partição no conjunto de dados analisado, encontra-se os seguintes 3 grupos de palavras (número definido pelo grupo):



Assim, como citado na seção anterior, 4 das palavras que ficaram agrupadas individualmente no dendograma anterior foram agrupadas em conjunto pelo k-means, especificamente no grupo 2. O grupo 1 é composto por 9 palavras um pouco menos abstratas e genéricas que as que encontram-se no grupo 2, havendo no caso termos como “Brasil” e “mercado”. Já no grupo 3, nota-se que conta com a maior quantidade de palavras, sendo que, de maneira geral, muitas delas não guardam muito significado por si só, como “assim”, “maior” e “neste”.

**Conclusão**

A partir do laboratório de análise de texto, foi possível observar a aplicação de algoritmos de inteligência artificial em dados não estruturados, no caso textos, os quais não são organizados em um formato fixo como tabelas com colunas e linhas definidas, bem como os desafios relacionados a trabalhar com eles, dados que esse tipo de dado em geral não é facilmente manipulado com técnicas tradicionais de machine learning e análise de dados, necessitando de uma etapa de pré-processamento mais complexa e aprofundada do que no caso do uso de dados estruturados, em geral.

Nesse sentido, o uso de gráficos e visualização como gráficos de frequência de palavras, nuvens de palavras, dendogramas e gráficos de cluster, facilita bastante o entendimento das análises para que seja possível gerar insights a partir disso de forma mais rápida, considerando a grande quantidade de dados textuais analisados no laboratório. A partir disso, é possível, por exemplo, identificar os principais tópicos abordados nos documentos e resumi-los de forma mais eficiente do que se o mesmo fosse feito de forma manual.

Além disso, é interessante analisar também a combinação de técnicas já estudadas em laboratórios anteriores, mas agora sendo aplicadas com dados textuais, como os agrupamentos hierárquicos e por partição. Entretanto, é importante pontuar nesse caso a relevância de preparar de forma adequada os dados antes de fazer as análises, buscando filtrar apenas termos relevantes para o estudo e não palavras sem tanto significado individualmente, como conectivos ou preposições. Esse foi um problema encontrado nos resultados do presente relatório, o que acabou não interferindo tanto no objetivo do laboratório, que era mais voltado para introduzir os alunos ao tema da análise de textos com machine learning aplicado a negócios, mas é válido pontuar como uma sugestão de melhoria para próximas aplicações.