

ACH-2018 - Inteligência Artificial

Trabalho (obrigatório) - Text Clustering

Profa. Dra. Sarajane Marques Peres

Thaís Rodrigues Neubauer e Caio Vinicius Canic Silva - bolsistas PAE

Universidade de São Paulo - Graduação em Sistemas de Informação

13 de março de 2018

Objetivo

Uma experimentação com os algoritmos de agrupamento (clustering): *K-means* e *Self Organizing Maps* - SOM. Estudo, implementação, testes e análise da aplicação desses algoritmos de agrupamento e dados do tipo texto. Essa tarefa é conhecida como *Text Clustering* ou *Text Document*.

Datas relevantes:

- 5 de junho: entrega do trabalho
- 19 de junho: teste de avaliação (individual)
- 26 de junho: início das apresentações

Entrega do trabalho e formação de grupos

- Todos os artefatos que compõem uma entrega, descritos nas próximas seções, deverão ser postados no e-Tidia. Uma atividade específica para cada entrega será publicada no e-Tidia.
- O horário de finalização do recebimento de arquivos é 23:55 - no horário do servidor no qual o Tidia está rodando.
- Atentem para a postagem na última hora: múltiplos usuários tentando submeter ao mesmo tempo podem causar congestão do servidor. Postagens ilimitadas serão possíveis, então **seja esperto** e suba versões do seu trabalho no decorrer do período de desenvolvimento do trabalho.
- Entregas após o horário e data limite poderão receber nota zero ou receber decréscimos de nota. Essa decisão é da professora da disciplina.
- **Para arquivos que sejam muito grandes e não possam ser colocados no Tidia, o grupo deve entregar em um pen-drive para a professora, antes do horário e data limite de entrega do trabalho. A professora não está presente na EACH 24/7, então de forma alguma deixem de se planejar com antecedência.**
- O grupo deve subir os arquivos na área de apenas um dos alunos do grupo.

Os alunos devem se organizar em grupos com no **máximo 5 pessoas**.

A avaliação do trabalho será realizada sobre cada um dos artefatos que compõem uma entrega: relatório, vídeo/código e apresentação.

- Eventualmente, durante a avaliação dos trabalhos, os grupos podem ser convocados para esclarecer algum aspecto da entrega submetida. Esse esclarecimento poderá ser realizado via entrega de arquivos adicionais, via conversa no Skype ou conversa pessoalmente - a depender do tipo do problema ou dúvida encontrada.
- Material de terceiros em domínio público (mantidos em sítios acadêmicos ou sítios especializados no assunto) poderão ser usados sob a condição de estarem claramente referenciados no relatório e de terem seu conteúdo analisado e compreendido por todos os membros do grupo.

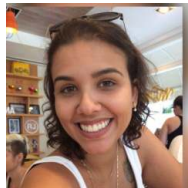
Cuidado com o plágio!!!!

Falhas em referenciar o uso de trabalho de terceiros caracteriza plágio. Se for constatado plágio de qualquer natureza durante a avaliação do trabalho, os integrantes do grupo receberão nota zero na entrega em questão.

Monitores da Disciplina



Caio Vinicius Canic Silva
Reconhecimento Biométrico Multimodal
baseado em Sinais de ECG, EEG e EMG



Thais Rodrigues Neubauer
Agrupamento Interativo em Mineração de Processos:
usando representações de dados baseadas
em frequência

Comunicação com os monitores

Toda comunicação com os monitores da disciplinas será feita via forum no ambiente Tidia. Não deixem para postar dúvidas na última hora!

Corpora

Um *corpus* é uma coleção de texto. Este trabalho versa sobre o agrupamento de textos que estão organizados em um corpus.

- Cada grupo deverá escolher dois *corpora* para executar o seu trabalho.
- Há corpora maiores e menores. Os grupos devem escolher pelo menos um que possua uma quantidade de textos maior (mais de 5000 textos). O segundo conjunto poder ser outro de grande porte ou um menor.
- A escolha pode ser feita dentre os *corpora* listados nesta especificação, ou pode vir de outros *corpora* disponíveis publicamente. **Para o último caso, o grupo deverá contatar os monitores e/ou a professora antes bater o martelo na escolha do corpus.**
- Alternativamente, o grupo pode criar um *corpus* (e neste caso, o segundo deve vir da lista fornecida), desde que o grupo siga regras específicas de: legislação de direitos autorais; características do corpus. **Se for esse o caso, o grupo deverá contatar os monitores e/ou a professora antes de iniciar o trabalho de geração de corpus.**

Lista de *corpus*:

- BBCsport <http://mlg.ucd.ie/datasets/bbc.html>, <http://mlg.ucd.ie/howmanytopics/index.html>
- BBC <http://mlg.ucd.ie/datasets/bbc.html>, <http://mlg.ucd.ie/howmanytopics/index.html>
- Newsgroups20: <https://www.kaggle.com/crawford/20-newsgroups>
- Process Mining Abstracts: disponível no Tidia
- Reuters-21578:
<https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection>
- Web-KB: <http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/>
- Guardian <http://mlg.ucd.ie/howmanytopics/index.html>
- Irishtimes <http://mlg.ucd.ie/howmanytopics/index.html>
- WikiLow <http://mlg.ucd.ie/howmanytopics/index.html>
- WikiHigh <http://mlg.ucd.ie/howmanytopics/index.html>

* em cinza: conjuntos que não disponibilizam os textos completos. Será mais complexo para fazer o pós-processamento.

Na tarefa de Text Clustering, os textos pertencentes ao corpus sob análise, serão processados pelo algoritmo de agrupamento (K-means ou SOM, no caso deste trabalho) a fim de descobrir quais devem ser indicados como pertencentes a um de vários grupos. Ao processar os textos, os algoritmos descobrirão características dos textos que os fazem similares mediante alguma perspectiva e, assim, formarão grupos de textos parecidos. Entretanto, os algoritmos fornecerão uma resposta sempre que forem executados, mas nem sempre será uma boa resposta. É tarefa sua trabalhar com os algoritmos de maneira que eles forneçam bons grupos - grupos que façam sentido.

Para alcançar esse objetivo, vocês precisarão:

- pré-processar os textos para que eles sejam representados de tal forma que os algoritmos sejam capazes de atuar sobre eles;
- pré-processar os textos para que eles formem um bom conjunto de dados para que os algoritmos sejam capazes de encontrar boas soluções para o agrupamento;
- implementar os algoritmos e explorar suas capacidades (parametrizar/calibrar) aplicando sobre os textos;
- aplicar medidas de qualidade de agrupamento e analisar os resultados dessas medidas;
- pós-processar e analisar os resultados de agrupamentos de maneira qualitativa (com semântica).

Para implementação das rotinas de pré-processamento você deve:

- estudar o relatório [PPgSI-001/2018](http://ppgsi.each.usp.br/relatorios-tecnicos-2018/) e explorar a nele bibliografia indicada:
<http://ppgsi.each.usp.br/relatorios-tecnicos-2018/>
- explorar a bibliografia disponibilizada no [repositório](#)
- implementar rotinas, ou aplicar rotinas prontas, de pré-processamento de textos e produzir uma série de conjuntos de dados para uso com os algoritmos K-means e SOM.

* Obviamente que quanto mais referências bibliográficas você acessar e ler, melhores condições você terá de produzir um bom trabalho.

Representações exigidas - requisito mínimo

- binária
- TF - *term frequency*
- TF-IDF - *term frequency - inverse document frequency*

** Toda e qualquer implementação de terceiros deve ser devidamente referenciada. Informe o uso de implementações de terceiros para esta fase no relatório e eventualmente no seu código, se trechos de terceiros forem usados dentro de uma codificação sua.

*** Para essa fase do trabalho, você pode usar implementações de código fechado, desde que exista documentação associada a ela que indique as estratégias que estão sendo usadas para realização de diferentes tarefas.

Implementação: K-means

Você vai implementar o K-means e suas variações!!!

A implementação do K-means, e suas variações, deverá ser de autoria dos alunos do grupo. Obviamente que o grupo tem liberdade de estudar implementações de terceiros, mas não deverá usá-las para execução do trabalho.

Implementações que deverão ser desenvolvidas:

- K-means clássico com inicialização de protótipos aleatória;
- K-means ++: K-means com inicialização especial;
- X-means: K-means com determinação automática do número de grupos.

Distâncias

Para o caso das implementações do K-means você deverá usar duas métricas: distância euclidiana e similaridade cosseno.

* É ~~preferível~~ necessário que o grupo utilize uma das seguintes linguagens de programação: R, Matlab, Python, Java, C ou C++. ~~Caso o grupo deseje utilizar outra linguagem é necessário antes entrar em contato com os monitores da disciplina para confirmar a viabilidade.~~

Você pode implementar o seu SOM ou usar implementações de terceiros

A implementação do SOM pode ser sua ou de terceiros. **Contudo**, sendo de terceiros:

- você deve ter certeza de que o código é aberto para todas as funções que implementam características do SOM;
- você deve ter certeza de que a implementação segue os princípios discutidos em aula;
- você deve ter completo conhecimento da lógica de programação utilizada para implementação cada uma das funções necessárias para aplicação do SOM.

Apenas a **distância euclidiana** será exigida para os testes com SOM, em termos de distância usada no espaço vetorial. A implementação com similaridade cosseno pode também ser feita, como uma característica extra do seu trabalho, sendo portanto opcional.

* É ~~preferível~~ necessário que o grupo utilize uma das seguintes linguagens de programação: R, Matlab, Python, Java, C ou C++. ~~Caso o grupo deseje utilizar outra linguagem é necessário antes entrar em contato com os monitores da disciplina para confirmar a viabilidade.~~

Implementações: medidas de desempenho

Todo resultado produzido por um algoritmo de agrupamento precisa ser analisado em termos de alguma medida de desempenho (quantitativo). A aplicação das medidas de avaliação sobre as execuções dos algoritmos gerarão uma série de números que deverão ser organizados em tabelas, ilustrado por meio de gráficos e discutidos no por meio de análises críticas (textuais). Algumas tabelas, gráficos e análises comporão o relatório e servirão como diretrizes para tomadas de decisões durante a execução do trabalho e para delineamento das conclusões.

Neste trabalho as seguintes medidas de desempenho deverão ser usadas:

- Para os resultados do K-means: [implementada pelo próprio grupo](#)
 - Silhouette
- Para os resultados do SOM: implementadas pelo próprio grupo ou disponíveis na implementação em uso
 - Erro de quantização
 - Erro de distorção topológica

Medidas extras podem ser aplicadas à escolha do grupo e se usadas além das medidas já citadas, valorizarão o trabalho do grupo.

O trabalho está localizado na área de análise de textos, mais especificamente, na descoberta de grupos, perfis ou tópicos. O resultado dos algoritmos de agrupamento precisam ser **explicados** para os interessados.

Os grupos deverão apresentar os resultados de agrupamentos finais de seu trabalho – os melhores resultados obtidos durante o desenvolvimento do trabalho – de maneira ilustrativa, fazendo uso de gráficos, histogramas, nuvens de palavras ou outros recursos que achar interessante.

Vejamos alguns exemplos. Os exemplos são apenas para mostrar formas de visualização. Não se trata de indicação de nenhuma bibliografia, ferramenta ou implementação.

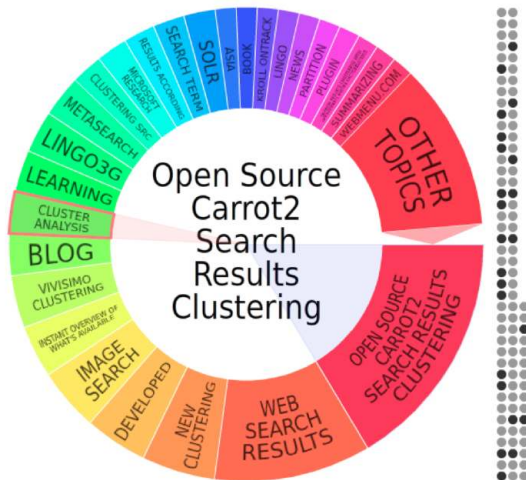


Figura 2: Exemplo 1: <https://en.wikipedia.org/wiki/Carrot2>

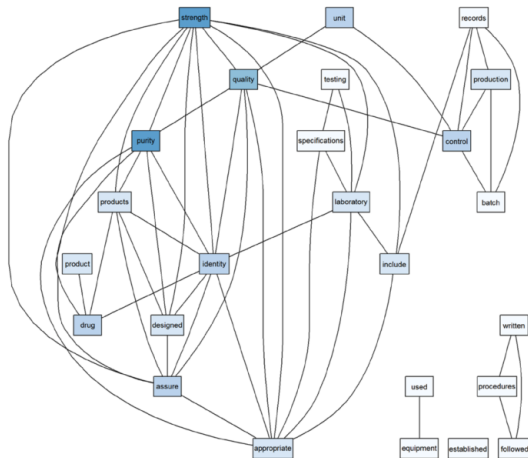


Figura 3: Exemplo 2:

<https://www.linkedin.com/pulse/fda-483-citations-text-mining-jose-i-rey>

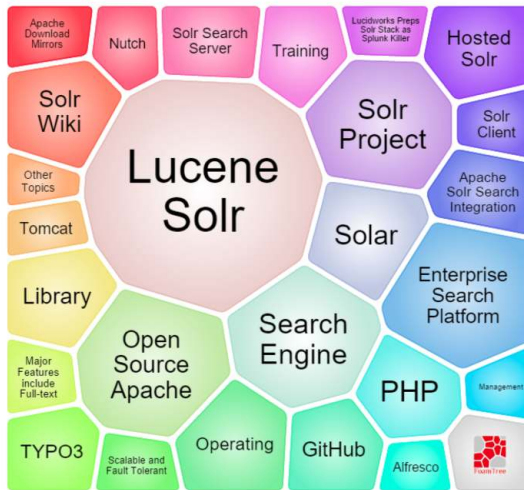


Figura 4: Exemplo 3:

https://lucene.apache.org/solr/guide/6_6/result-clustering.html

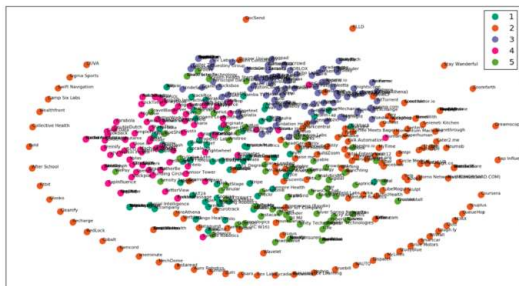


Figura 5: Exemplo 4: <https://stackoverflow.com/questions/37532040/interpreting-cluster-results-on-text?rq=1>

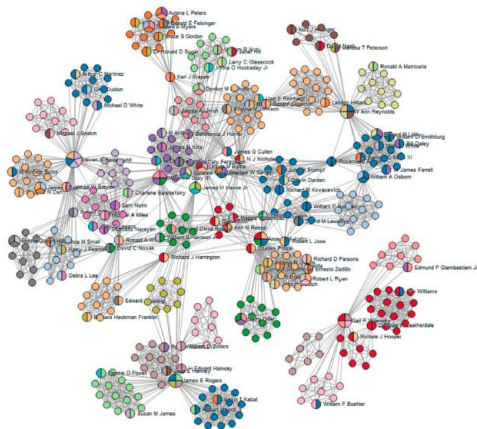


Figura 6: Exemplo 5:

<https://bulaza.wordpress.com/2011/11/29/orgpedia-board-members-network/>

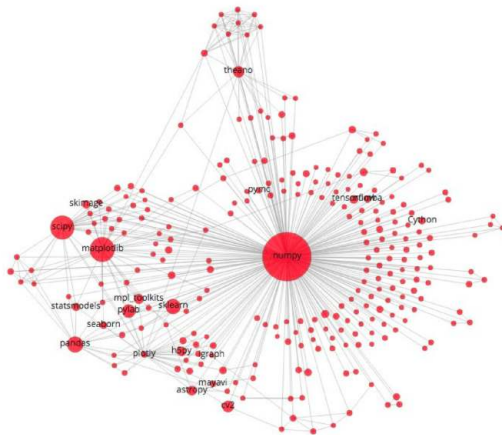


Figura 7: Exemplo 6: <https://twitter.com/plotlygraphs/status/752978082617667584>

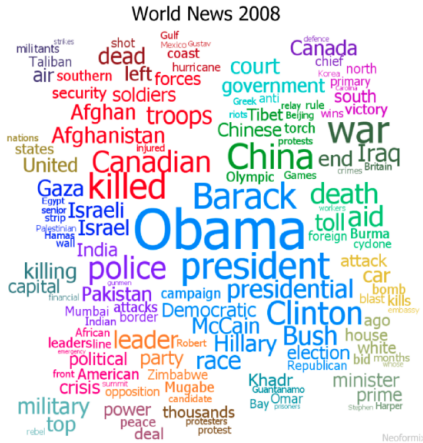


Figura 8: Exemplo 7:

<https://neoformix.com/2009/WorldNewsClusteredWordCloud.html>

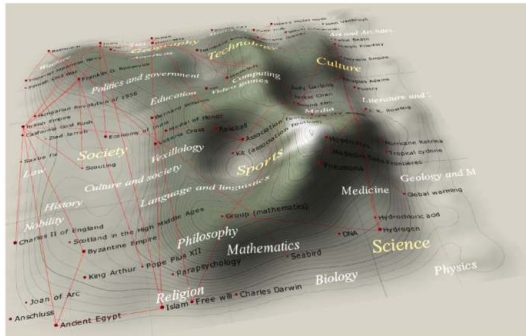


Figura 9: Exemplo 8: https://en.wikipedia.org/wiki/Self-organizing_map

Pós-processamento

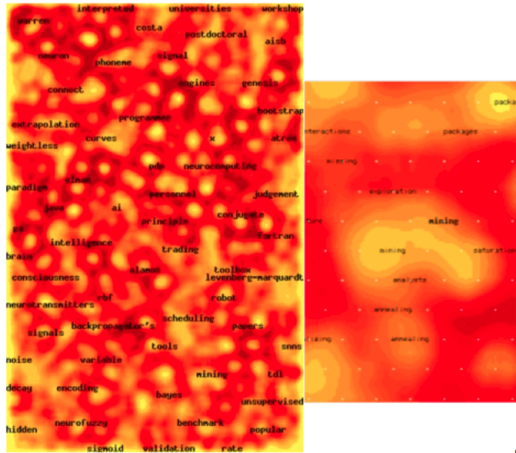


Figura 10: Exemplo 9: <http://slideplayer.com/slide/10890014/>

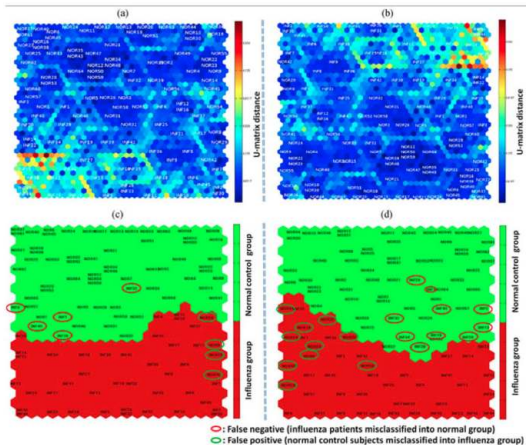


Figura 11: Exemplo 10: http://file.scirp.org/Html/2-8202392_36063.htm

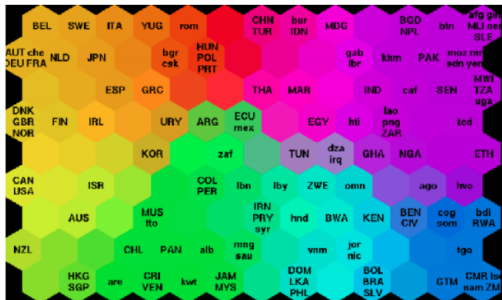


Figura 12: Exemplo 11:

<https://stackoverflow.com/questions/43257141/mapping-data-to-an-som-in-r>

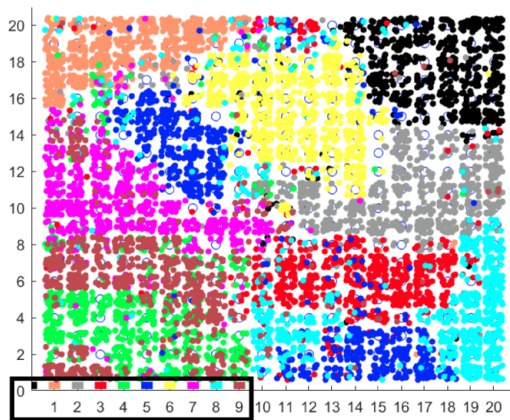


Figura 13: Exemplo 12: <http://blog.yhat.com/posts/self-organizing-maps-2.html>

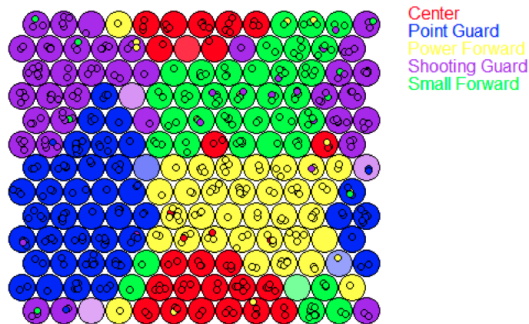


Figura 14: Exemplo 13: <https://digitalprojectstudio.wordpress.com/2017/04/13/introduction-to-self-organizing-maps-in-r/>

- Comente **detalhadamente** as linhas de código referentes às lógicas dos algoritmos. Seus comentários devem possibilitar que nós possamos encontrar facilmente os trechos de código das principais características dos algoritmos.
- Prepare seu código para salvar a progressão do algoritmo, em termos de medidas de desempenho, durante o treinamento.
- Prepare seu código para criar arquivos de parametrização, de forma que seja possível associar arquivos de resultados com arquivos de parametrização que geraram os resultados.
- Prepare arquivos detalhados do tipo Readme.txt explicando do que trata cada arquivo entregue e como executar o código.

Objetivo

O objetivo do vídeo é apresentar rapidamente as características da codificação construída. **Todos os itens devem ser explicados usando o código-fonte do seu trabalho.**

Requisitos para o vídeo:

- Duração de 10 a 20 min, formato MP4, resolução suficiente para o código estar legível e cadência normal para o discurso (fala). [Não acelere o vídeo.](#)
- Para o K-means, apresentar: estruturas de dados que organizam os protótipos do K-means e suas variações; implementação do cálculo da distância euclidiana e similaridade cosseno; implementação das estratégias de inicialização; implementação da estratégia de determinação do número de grupos, implementação de critérios de parada.
- Para o SOM, apresentar: estruturas de dados para organização dos espaços vetorial e matricial; implementação do algoritmo de aprendizado incluindo controle de taxa de aprendizado, função e raio de vizinhança, alteração de pesos; implementação de critérios de parada.
- Para as medidas de desempenho: apresentar a implementação de cada uma delas (Silhoutte, erro de quantização e erro de distorção topológica).

- O relatório deverá ser elaborado seguindo o formato IEEE, na opção “*Template for Transactions*”. As seções sugeridas não precisam ser seguidas: a ideia é usar a mesma diagramação, tamanho e tipo de fonte, estilo dos parágrafos, margens, referências bibliográficas, etc. O arquivo deve ser convertido no formato PDF antes da submissão da entrega.
- Recomendamos fortemente o uso do editor \LaTeX .
- O relatório NÃO DEVE CONTER seções explicando a teoria dos algoritmos usados e nem tampouco teoria sobre *text clustering*, ou outras teorias.
- O relatório deve ser organizados nas seguintes partes:
 - **Preliminares:** seção na qual o grupo tem a liberdade de trazer informações gerais sobre o ambiente de programação usado, particularidades para execução de seu código, problemas encontrados e como foram resolvidos; referências para códigos de terceiros usados no decorrer do trabalho;
 - **Conjuntos de dados:** informe os conjuntos de dados escolhidos e se recortes nos conjuntos foram feitos, dê todas as informações necessárias para reprodução desses recortes. Ainda com essa explicação, todos os recortes realizados devem ser mantidos em poder do grupo até que a nota final do trabalho seja publicada. No caso de contestação de nota, o grupo deverá necessariamente entregar todos os recortes usados no trabalho para a professora.

- continuação
 - **Seção de pré-processamento:** relate, em detalhes, todas as tomadas de decisão realizadas durante a fase de pré-processamento de dados. Motive suas decisões. Informe o grau de esparsidade das matrizes de dados geradas em cada uma das representações para cada um dos conjuntos de dados (e seus recortes) utilizados no trabalho. Crie uma tabela para organizar esses dados. Gráficos podem também ser usados em adição às tabelas.
 - **Exploração do K-means e pós-processamento de seus resultados:** as estratégias de aplicação do K-means e suas variantes deverá ser detalhadamente apresentadas. CUIDADO: organize sua apresentação. De todas as execuções realizadas, algumas deverão ser escolhidas para ter seus resultados apresentados no trabalho. Você deve escolher pelo menos uma para cada variação de K-means apresentada. Descreva a sua estratégia de parametrização dos algoritmos e de escolha dos resultados a serem apresentados. As medidas de qualidade de agrupamento e visualizações semânticas discutidas neste documento deverão ser discutidas e ilustradas no seu relatório.

- continuação
 - **Exploração do SOM e pós-processamento de seus resultados:** as estratégias de aplicação do SOM. De todas as execuções realizadas, algumas deverão ser escolhidas para ter seus resultados apresentados no trabalho. Você deve escolher pelo menos três execuções do SOM para discutir. Descreva a sua estratégia de parametrização dos algoritmos e de escolha dos resultados a serem apresentados. As medidas de qualidade de agrupamento e visualizações semânticas discutidas neste documento deverão ser discutidas e ilustradas no seu relatório.
 - **Conclusões:** você deve discorrer sobre suas impressões sobre os algoritmos, comparando-os, levantando vantagens e desvantagens, discutindo seus desempenhos na tarefa de *text clustering*.

Teste avaliativo e apresentação do trabalho

Teste avaliativo

O teste avaliativo será uma prova escrita, individual, na qual serão feitas perguntas sobre o trabalho. Essas perguntas podem versar sobre qualquer aspecto do trabalho, desde detalhes da teoria relacionada aos aspectos dos algoritmos que foram implementados, até detalhes sobre estratégias e tomadas de decisão realizadas no trabalho e resultados obtidos.

Apresentação do trabalho

O formato da apresentação será fornecido depois da entrega dos trabalhos, pois precisamos saber quantos grupos entregarão o trabalho. Mas já é possível informar que todos os alunos precisam estar presentes na apresentação para receber a nota referente a ela e precisam participar da apresentação. O principal assunto da apresentação será sobre os resultados obtidos.

Esclarecendo dúvidas:

- Quando você escolhe um corpus para atender ao requisito do trabalho sobre o uso de um corpus com mais de 5000 textos, você não necessariamente precisa usar o conjunto de dados escolhido em sua completude. Você pode, por exemplo, amostrar o conjunto e usar 5000, 10000 ou 150000 dados - conforme você evolui no trabalho você aumenta a complexidade do seu conjunto de dados.



Profa. Dra. Sarajane Marques Peres
Universidade de São Paulo
Escola de Artes, Ciências e Humanidades
Sala 320-A - Bloco I1
sarajane@usp.br
www.each.usp.br/sarajane