projeto

November 22, 2019

1 Projeto Final

Este é o projeto final da disciplina Aprendizado de Máquina (IA006-C), ministrado pelos professores Levy Boccato e Romis, na Unicamp no 2S2019.

1.1 Projeto

A ideia do projeto é permitir a clusterização de conteúdo textual, para que a partir deste seja criado um chatbot.

Os textos passarão pro um processo de clusterização (e aqui serão apresentados duas técnicas para gerar o espaço vetorial de documentos [TF-IDF e Doc2Vec]) usando o algoritmo KMeans e usando duas métricas para cálculo das distâncias dos documentos no espaço vetorial desejado.

Posterior a isso, textos que não forem similares (ou proximamente similares aos já "classificados") serão considerados como anomalias e por conseguintes novos clusters poderão ser gerados futuramente.

1.1.1 Carregamento dos datasets

Os datasets de exemplos são frases já pré-categorizadas usadas em chatbots.

Contém 32 categorias e ao todo 690 documentos ou frases.

```
<IPython.core.display.HTML object>


Qtde. de documentos por categoria:


<IPython.core.display.HTML object>


Total docs     : 690
Total cluster  : 690
X_train size   : (552,)
X_test  size   : (138,)
```

1.1.2 Dataset tokenization

```
Tokenization...
Qtd documentos treino:  552
Qtd Intents treino   :  32
Finished...
```

```
Out[7]: [TaggedDocument(words=['ir', 'estar', 'viajar', 'certar', 'empresar', 'fornecer', 'cobe
         TaggedDocument(words=['qual', 'outro', 'servico', 'bot', 'oferecer'], tags=[1]),
         TaggedDocument(words=['um', 'nome'], tags=[2]),
         TaggedDocument(words=['nao', 'poder', 'recarregar', 'porque', 'dizer', 'sim', 'desativ
         TaggedDocument(words=['poder', 'ajudar', 'mudar', 'tocar', 'telefonar'], tags=[4]),
         TaggedDocument(words=['configurar', 'email', 'thunderbird'], tags=[5]),
         TaggedDocument(words=['nao', 'precisar', 'mais', 'nado', 'obrigar'], tags=[6]),
         TaggedDocument(words=['nao', 'conseguir', 'acessar', 'web', 'telefonar', 'nao', 'certe
         TaggedDocument(words=['poder', 'ter', 'correar', 'voz', 'permanentemente', 'remover',
         TaggedDocument(words=['tutorial', 'configuracao', 'redar', 'fiar', 'android'], tags=[9
```

1.1.3 Doc2Vec

Parâmetros iniciais... quantidade de dimensões dos vetores gerados para cada frase, épocas de treinamento e épocas de posterior inferência para novas frases.

A quantidade de épocas de inferência, sugere-se ser bem superior as de treinamento.

```
Dimensions   : 1500
Epochs       : 200
Infer Epochs : 15000
```

```
Starting model...
Building vocab...
Training...
Finish...
```

Validação do modelo gerado pelo Doc2Vec... teste tanto nos dados apresentados para treinamento quanto nos dados de testes e as acurácias alcançadas.

Randomicamente escolhendo 100 amostras de teste.

```
- Acurácia treino: 100.0
 - Acurácia teste 1  : 54.0
 - Acurácia teste 2  : 54.0
 - Acurácia teste 3  : 56.0
 - Acurácia teste 4  : 55.0
 - Acurácia teste 5  : 56.0
- Acurácia média teste: 55.0
```

Clusterização Utilizou-se o KMeans definindo a quantidade de clusters para o número ideal de categorias existentes no caso 33. A métrica de distância utilizada, não foi a euclidiana, mas sim a de cosseno (métrica comumente usada na classificação de texto em seu espaço vetorial).

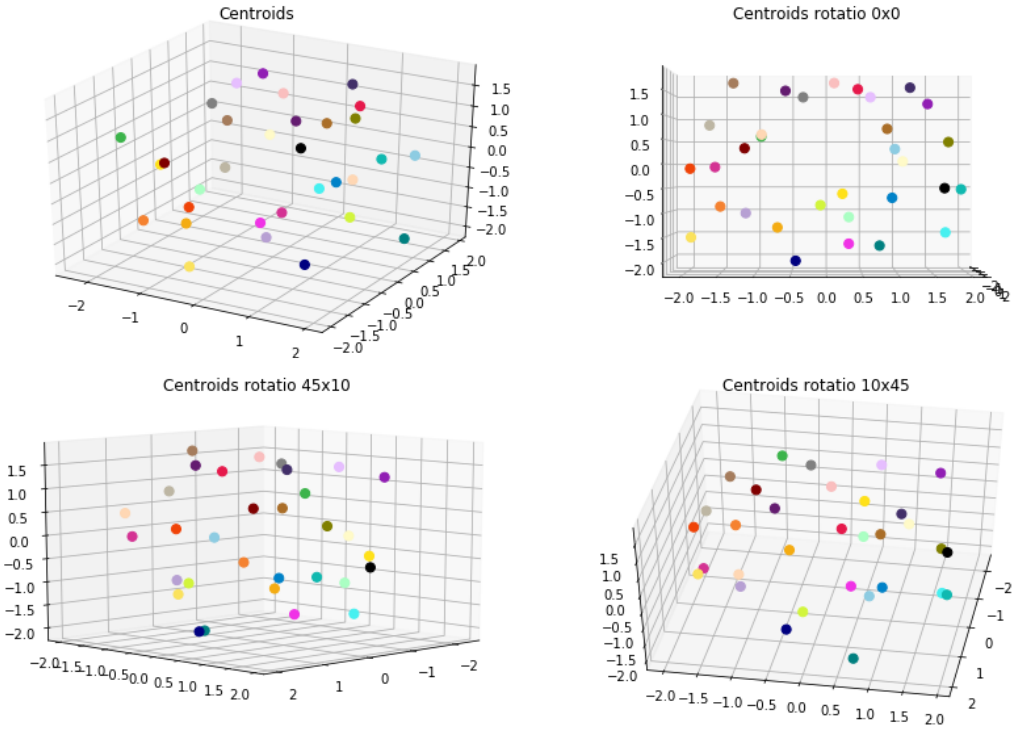
```
Frases por cluster:


<IPython.core.display.HTML object>


--------------------
Documentos por cluster:


<IPython.core.display.HTML object>
```

Visualização Apresentação dos protótipos gerados pelo KMeans, reduzindo a dimensão usando o algoritmo MDS (Multidimensional Scaling).



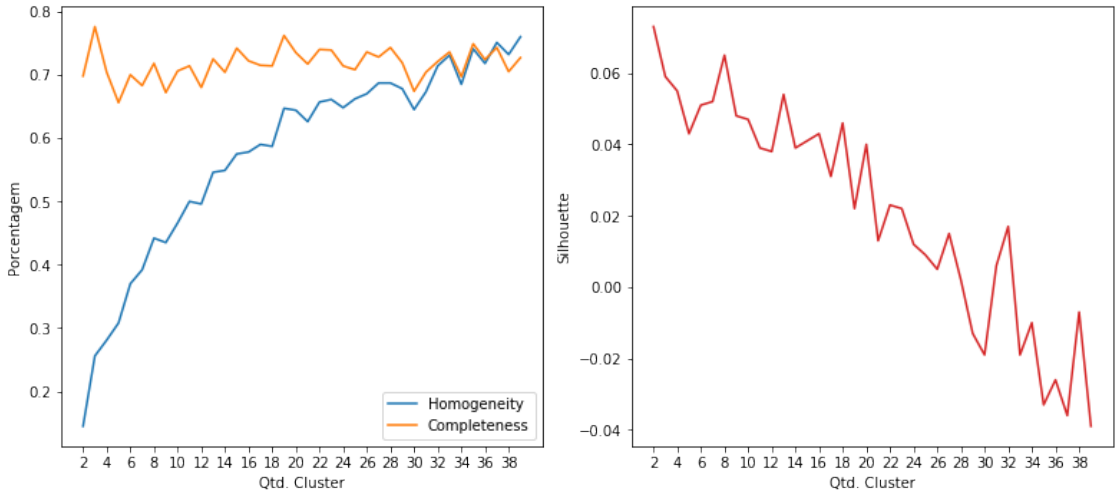
Clusterização dos dados de Teste Por fim, realizada a clusterização dos dados de teste e a apresentação das 8 primeiras frases do conjunto de teste juntamente com outras duas frases do cluster ao qual foi identificado como o melhor.

```
* Se eu ativar roaming agora, já posso utilizar?
  - Eu quero ativar o identificador de chamadas.
  - Se eu cancelar o roaming agora, demora muito pra concluir o pedido?
--------------------
* Fui assaltado, preciso acionar seguro.
  - Oi, eu preciso de um seguro para o meu telefone, quais são as opções disponíveis para o me
  - Tom, quais são as políticas para retornar um dispositivo que foi usado por um tempo?
--------------------
* Estou bastante cansado do meu telefone existente e gostaria de saber quando eu vou ser elegív
  - Eu não posso recarregar porque diz que meu sim está desativado , meu sim pode ser ativado?
  - Por que sua loja não aceita o meu retorno?
--------------------
* Oi, como posso ativar o encaminhamento de chamada no meu plano?
  - Estou recebendo chamadas indesejadas de números desconhecidos, como posso ativar Não pertu
  - Teria sido muito difícil para me alertarem que uma mudança de planos me faria perder a mir
--------------------
* Eu não consigo acessar meu email
  - como alterar minha senha do usuário?
  - Não consigo recuperar a minha senha
--------------------
* Estou indo para uma viagem ao extremo norte da Ilha de Vancouver e gostaria de verificar se h
  - Como faço chamadas telefônicas para outros países?
  - posso acessar meu endereço eletrônico via thunderbird ou outlook?
--------------------
* Gostaria de verificar todas as taxas para cobertura internacional
  - Eu vou à Florença neste fim de semana e quero ativar roaming.
  - Olá quero verificar meu status de roaming internacional
--------------------
* Além do fornecimento de água vocês oferecem algum outro serviço?
  - tem algum outra opção de atendimento?
  - essas opções não tem o que estou procurando
--------------------
```

Métricas Abaixo são apresentadas métricas para demonstrar o quanto a clusterização parece funcionar.

```
Homogeneidade    :  0.701
Completude       :  0.729
Silhouette       :  0.017
```

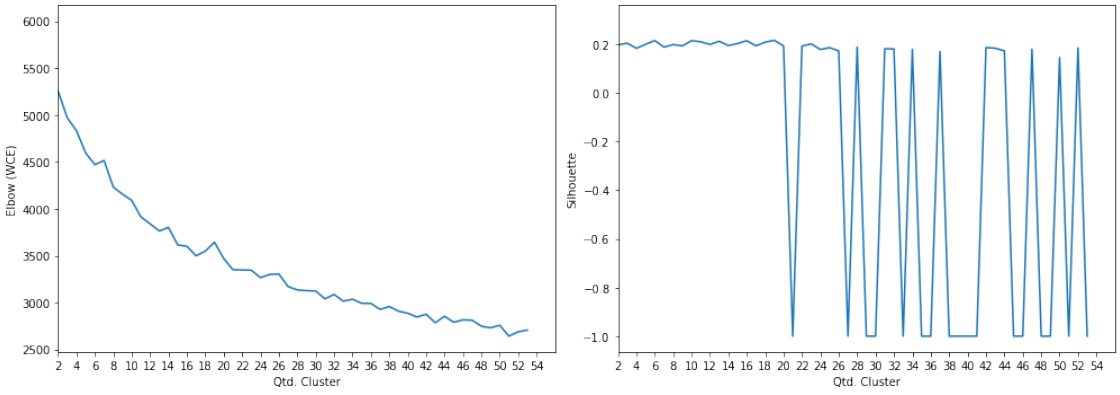
Como exemplo de comparação, foi executado o mesmo algoritmo de clusterização (conforme apresentado acima) entretanto variando a quantidade do número de clusters para verificar como as métricas se comportam.



Escolha da quantidade de Cluster Como não sabe-se ao certo quantos clusteres na realidade podem vir a existir, considerou-se que a quantidade máxima de clusters seria algo em torno de 10% da quantidade de dados existentes.

Para calcular exatamente qual a quantidade máxima, utilizou-se do maior valor dados pelas métricas Elbow e Silhouette (cada uma dando seu valor ideal de clusteres).

Abaixo segue o resultado.




```
Frases por cluster:


<IPython.core.display.HTML object>


--------------------
Documentos por cluster:
```

```
Out[21]: <IPython.core.display.HTML object>
```

1.1.4 TF-IDF

No caso do tf-idf, assim como no doc2vec foi escolhido um máximo de até 1500 features (ou dimensões). Entretanto, diferentemente do doc2vec o tf-idf não adiciona dimensões caso a quantidade de termos (palavras) seja inferior a esse máximo, mas ele corta caso for maior.

```
Tokenization...
Qtd documentos treino:  552
Qtd Intents treino   :  32
Finished...
```

```
Out[23]: ['ir estar viajar certar empresar fornecer cobertura telefonar area viagem poder dar
          'qual outro servico bot oferecer',
          'um nome',
          'nao poder recarregar porque dizer sim desativado sim poder ser ativado',
          'poder ajudar mudar tocar telefonar',
          'configurar email thunderbird',
          'nao precisar mais nado obrigar',
          'nao conseguir acessar web telefonar nao certeza precisar algum configuracao',
          'poder ter correar voz permanentemente remover planar',
          'tutorial configuracao redar fiar android']
```

Validação do modelo gerado pelo TF-IDF... teste tanto nos dados apresentados para treinamento quanto nos dados de testes e as acurácias alcançadas.

Randomicamente escolhendo 100 amostras de teste.

```
- Acurácia treino: 99.0
--------------------
 - Acurácia teste 1  : 50.0
 - Acurácia teste 2  : 57.0
 - Acurácia teste 3  : 55.0
 - Acurácia teste 4  : 57.0
 - Acurácia teste 5  : 57.0
- Acurácia média teste: 55.2
```

Clusterização Utilizou-se o KMeans definindo a quantidade de clusters para o número ideal de categorias existentes no caso 33. A métrica de distância utilizada, não foi a euclidiana, mas sim a de cosseno (métrica comumente usada na classificação de texto em seu espaço vetorial).

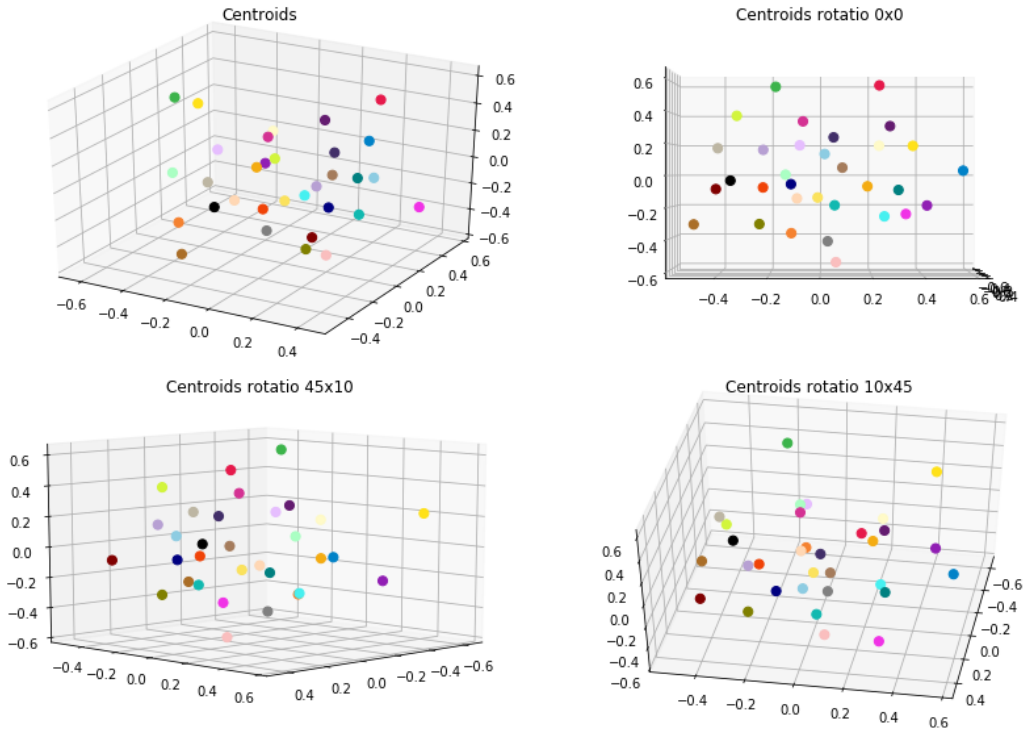
```
Frases por cluster:
```

```
<IPython.core.display.HTML object>
```

```
--------------------
Documentos por cluster:


Out[28]: <IPython.core.display.HTML object>
```

Visualização Apresentação dos protótipos gerados pelo KMeans, reduzindo a dimensão usando o algoritmo MDS (Multidimensional Scaling).



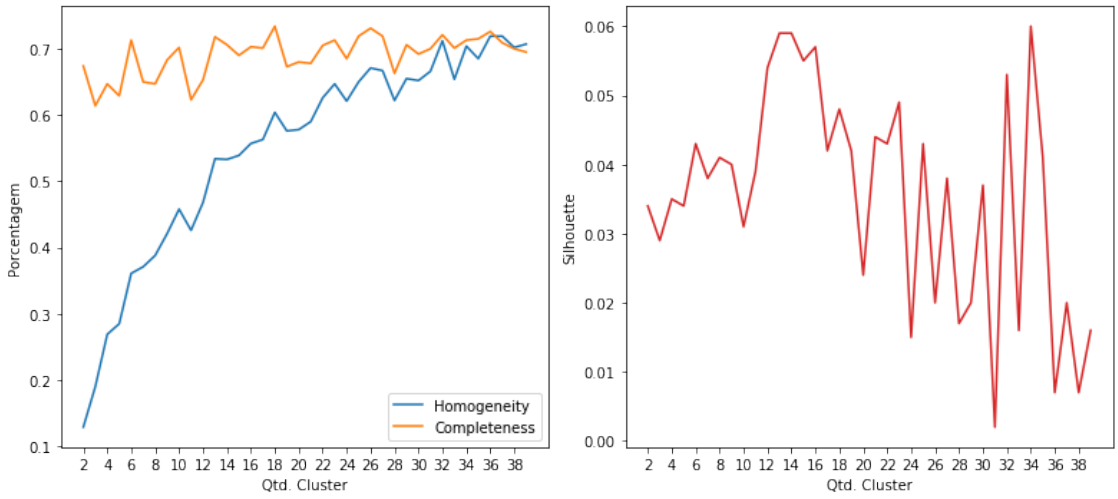
Clusterização dos dados de Teste Por fim, realizada a clusterização dos dados de teste e a apresentação das 8 primeiras frases do conjunto de teste juntamente com outras duas frases do cluster ao qual foi identificado como o melhor.

```
* Se eu ativar roaming agora, já posso utilizar?
  - Ajuda com a desativação de roaming
  - Desativar meu dispositivo quando no exterior
--------------------
* Fui assaltado, preciso acionar seguro.
  - preciso de ajudar
  - Posso trocar meu equipamento para o mais recente e o melhor disponível?
--------------------
* Estou bastante cansado do meu telefone existente e gostaria de saber quando eu vou ser elegí
  - Por que sua loja não aceita o meu retorno?
```

```
  - Posso cancelar meu plano pré-pago a qualquer momento?
--------------------
* Oi, como posso ativar o encaminhamento de chamada no meu plano?
  - Qual plano de dados é o menos caro para mim?
  - Qual é o mais recente plano 4G adequado para mim como um pacote família?
--------------------
* Eu não consigo acessar meu email
  - Como faço para recuperar meus emails?
  - consigo recuperar emails perdidos?
--------------------
* Estou indo para uma viagem ao extremo norte da Ilha de Vancouver e gostaria de verificar se
  - O meu telefone não recebe cobertura quando eu dirijo para a casa da minha mãe. Se eu lhe
  - Onde estão as áreas perto de mim que eu poderia ter boa recepção?
--------------------
* Gostaria de verificar todas as taxas para cobertura internacional
  - Ligue a função de roaming internacional.
  - Como posso ativar o roaming internacional, por favor?
--------------------
* Além do fornecimento de água vocês oferecem algum outro serviço?
  - você atende outro tipo de serviço? ou só estes?
  - você tem outros serviços?
--------------------
```

Métricas Abaixo são apresentadas métricas para demonstrar o quanto a clusterização parece funcionar.

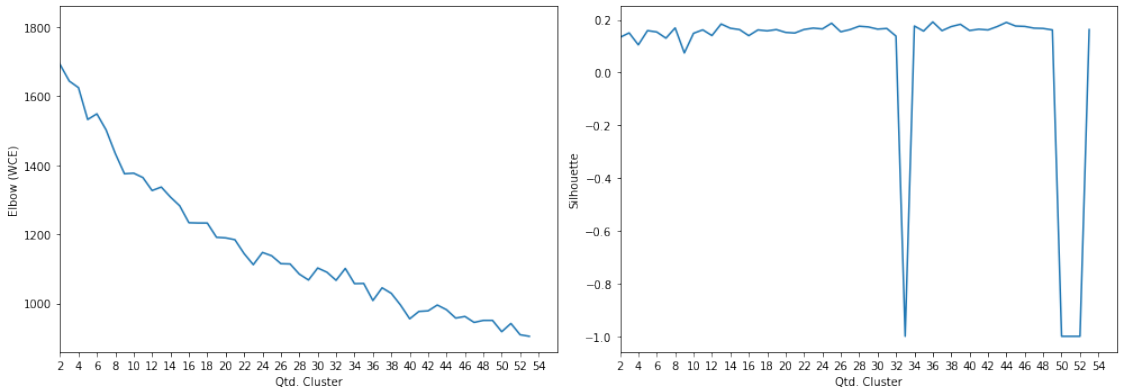
```
Homogeneidade:  0.704
Completude   :  0.731
Silhouette   :  0.025
```



Escolha da quantidade de Cluster Como não sabe-se ao certo quantos clusteres na realidade podem vir a existir, considerou-se que a quantidade máxima de clusters seria algo em torno de 10% da quantidade de dados existentes.

Para calcular exatamente qual a quantidade máxima, utilizou-se do maior valor dados pelas métricas Elbow e Silhouette (cada uma dando seu valor ideal de clusteres).

Abaixo segue o resultado.




```
Frases por cluster:


<IPython.core.display.HTML object>


--------------------
Documentos por cluster:


Out[36]: <IPython.core.display.HTML object>
```