# projeto

December 1, 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.

| | perguntas | cluster |
|---|---|---|
| 198 | como faco para trocar o meu usuario | ACCOUNT |
| 204 | eu consigo trocar meu username para outro? | ACCOUNT |
| 237 | estou falando com um bot não? | BOT_FOUND |
| 226 | quero solicitar a renovação de um certificado digital | CERTIFICATE |
| 148 | estou sem acesso ao meu endereco eletronico | EMAIL |
| 167 | como saber mais sobre o email da empresa | EMAIL |
| 113 | quais outras opções tem para me mostrar? | NO_OPTION |
| 119 | nenhuma dessas opções me ajuda | NO_OPTION |
| 23 | Não consigo trocar a minha senha | PASSWORD |
| 0 | posso consultar informações relativas a outros serviços e projetos? | SERVICES |

Qtde. de documentos por categoria:

<IPython.core.display.HTML object>

Total docs : 272
Total cluster : 272
X_train size : (217,)
X_test size : (55,)

### 1.1.2 Dataset tokenization

Tokenization...
Qtd documentos treino: 217
Qtd Intents treino : 12
Finished...

Out[8]: [TaggedDocument(words=['existir', 'algum', 'maneirar', 'alterar', 'nome', 'usuario'], tags=[0]),
TaggedDocument(words=['nao', 'precisar', 'mais'], tags=[1]),
TaggedDocument(words=['configurar', 'outlook'], tags=[2]),
TaggedDocument(words=['senha', 'acessar'], tags=[3]),
TaggedDocument(words=['certificar', 'digitar'], tags=[4]),
TaggedDocument(words=['mais', 'email', 'empresar'], tags=[5]),
TaggedDocument(words=['necessario', 'instalar', 'algum', 'software', 'adicional', 'conectar', 'redar', 'fiar'], t
TaggedDocument(words=['alterar', 'senha', 'usuario'], tags=[7]),
TaggedDocument(words=['opcao', 'ajudar'], tags=[8]),
TaggedDocument(words=['email', 'nao', 'entrar', 'acessar'], 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 : 1000
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: 99.0
 - Acurácia teste 1  : 80.0
 - Acurácia teste 2  : 80.0
 - Acurácia teste 3  : 80.0
 - Acurácia teste 4  : 80.0
 - Acurácia teste 5  : 80.0
- Acurácia média teste: 80.0

**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).

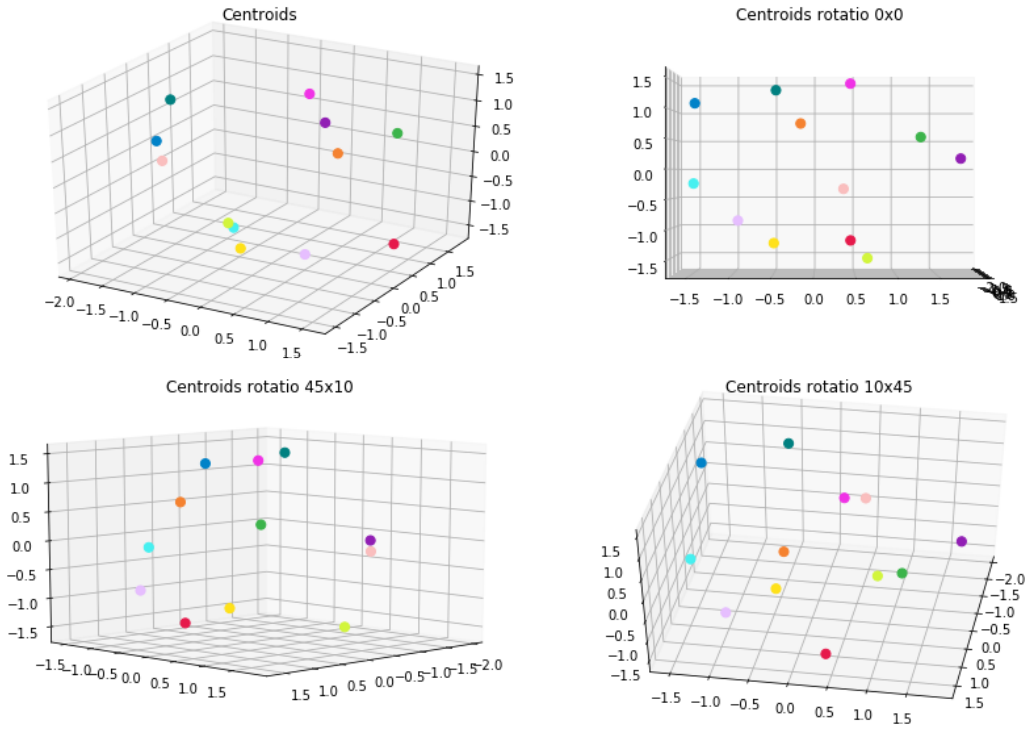
Frases por cluster:

<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).



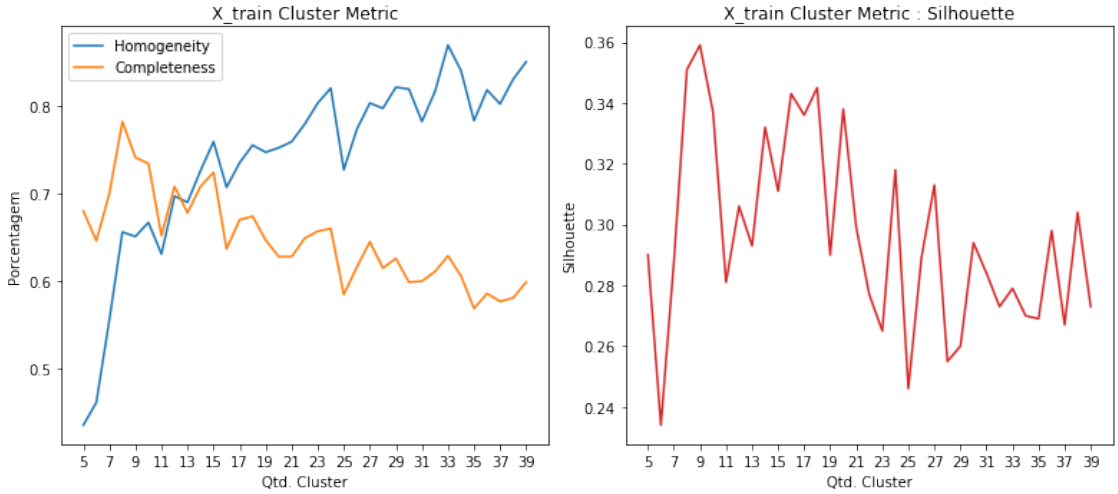
**Clusterização dos dados de Teste** Por fim, realizada a clusterização dos dados de teste e a apresentação das 8 primeiras frases do conjunto de teste juntamente com outras duas frases do cluster ao qual foi identificado como o melhor.

* Meu email está com problema
  - Meu email não entra, sem acesso
  - Como eu configuro meu email no thunderbird
---
* existem mais serviços com os quais eu posso consultar?
  - quero consultar outro tipo de informação com você, posso?
  - que tipos de serviços você oferece?
---
* como eu faco para criar um novo usuario?
  - é permitido que patrulheiros tenham uma conta?
  - quero criar uma conta de usuário para um colaborador externo, como fazer?
---
* como me conectar ao wifi (rede sem fio) da empresa
  - e necessario instalar algum software adicional para conectar na rede sem fio?
  - preciso conectar me a internet através da rede sem fio (wifi)
---
* eu consigo alterar meu nome de usuário para outro?
  - não estou conseguindo criar um novo nome de usuário
  - tem como trocar o meu username?
---
* qual usuario e senha usar para acessar os sistemas?
  - existe alguma maneira de alterar meu nome de usuário?
  - tem como eu mudar meu nome de usuário?
---
* Não consigo entrar no meu email
  - estou sem acesso ao meu email
  - Meu email não entra, sem acesso
---
* tem como renovar um certificado digital emitido?
  - certificados digitais
  - Revogação de certificado digital
---

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

Homogeneidade : 0.79
Completude : 0.804
Silhouette : 0.137

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.

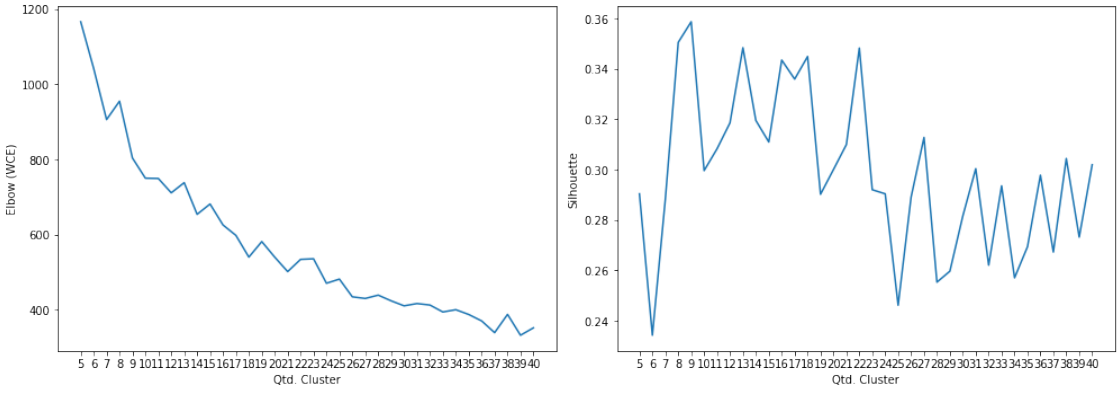



**Escolha da quantidade de Cluster** Como não sabe-se ao certo quantos clusteres na realidade podem vir a existir, considerou-se que a quantidade máxima de clusters seria algo em torno de 40.

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.

```
Running Elbow...
Running Silhouette...
N. Elbow      Cluster :  18
N. Silhouette Cluster :  9
```



Dados treinamento
---
Homogeneity : 61.0
Completeness : 70.0
V-Measure : 65.0
Silhouette : 0.30134716629981995

Dados teste
---
Homogeneity : 69.0
Completeness : 79.0
V-Measure : 74.0
Silhouette : 0.11756753921508789

Frases por cluster:

<IPython.core.display.HTML object>

---
Documentos por cluster:

Out[28]: <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 500 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: 217
Qtd Intents treino : 12
Finished...

Out[30]: ['existir algum maneirar alterar nome usuario',
'nao precisar mais',
'configurar outlook',
'senha acessar',
'certificar digitar',
'mais email empresar',
'necessario instalar algum software adicional conectar redar fiar',
'alterar senha usuario',
'opcao ajudar',
'email nao entrar acessar']

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: 87.0
---
- Acurácia teste 1 : 78.18
- Acurácia teste 2 : 78.18
- Acurácia teste 3 : 78.18
- Acurácia teste 4 : 78.18
- Acurácia teste 5 : 78.18
- Acurácia média teste: 78.18

**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).

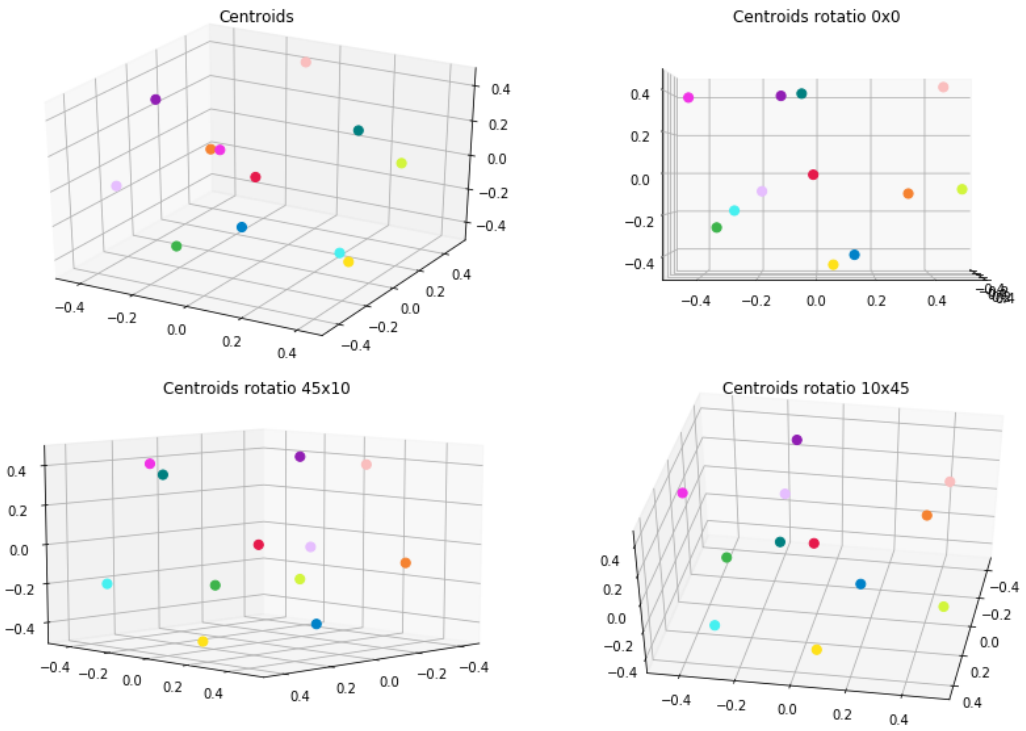
Frases por cluster:

<IPython.core.display.HTML object>

---
Documentos por cluster:

Out[35]: <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.

* Meu email está com problema
  - vc tem nome?
  - Como redirecionar meus emails para outro endereço

---

* existem mais serviços com os quais eu posso consultar?
  - existem outros programas para acessar meu email?
  - posso consultar informações relativas a outros serviços e projetos?

---

* como eu faco para criar um novo usuario?
  - como trocar minha senha
  - tem como trocar o meu nome de usuário para outro?

---

* como me conectar ao wifi (rede sem fio) da empresa
  - quero me conectar a rede sem fio
  - como me conectar a rede sem fio

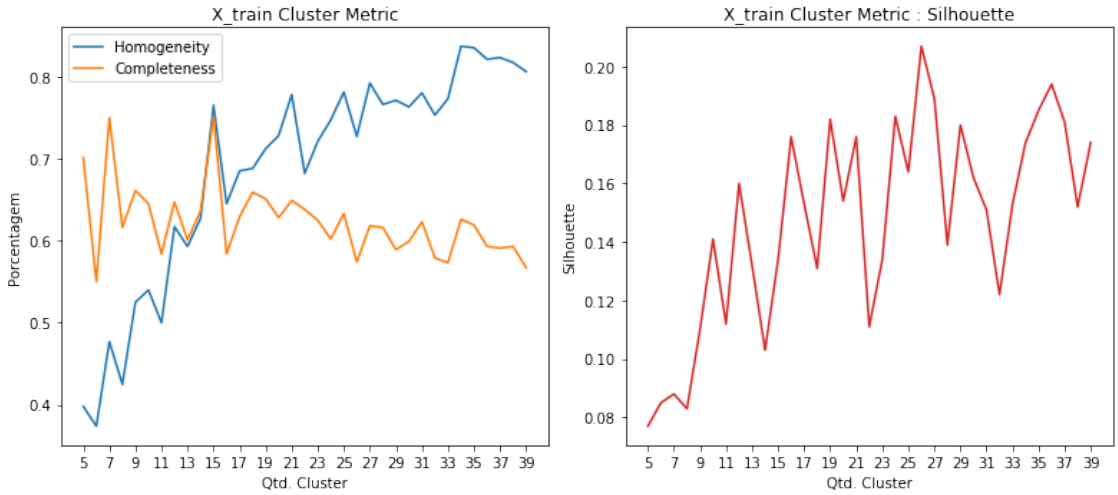
---

* eu consigo alterar meu nome de usuário para outro?
  - como trocar minha senha
  - tem como trocar o meu nome de usuário para outro?

---
* qual usuario e senha usar para acessar os sistemas?
  - como faco para acessar a rede sem fio de fora da empresa?
  - quero acessar a rede da empresa da minha casa
---
* Não consigo entrar no meu email
  - o email não esta no spam
  - estou tentanto criar meu username, mas não estou conseguindo
---
* tem como renovar um certificado digital emitido?
  - como emitir novos certificados digitais?
  - como faço para criar um novo certificado digital?
---

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

Homogeneidade: 0.773
Completude : 0.742
Silhouette : 0.191



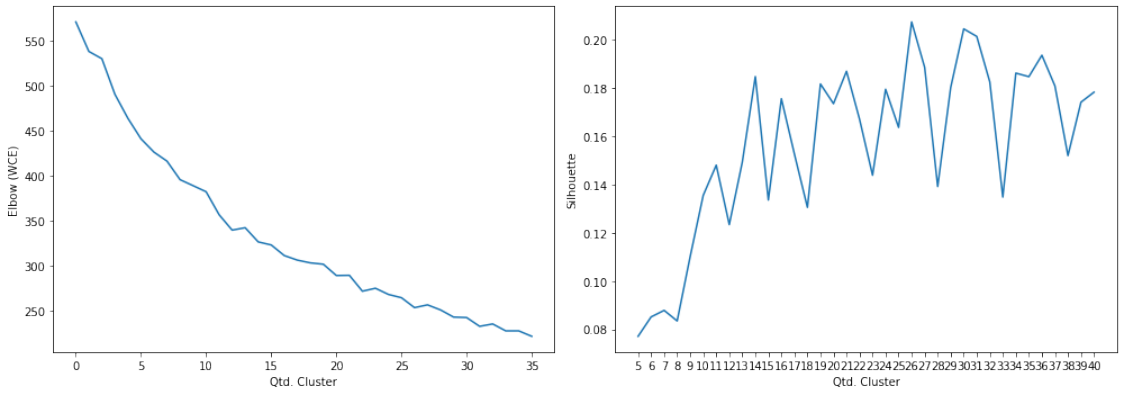

**Escolha da quantidade de Cluster** Como não sabe-se ao certo quantos clusteres na realidade podem vir a existir, considerou-se que a quantidade máxima de clusters seria algo em torno de 40.

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.

Running Elbow...
Running Silhouette...

N. Elbow Cluster : 18
N. Silhouette Cluster : 26




Dados treinamento
---
Homogeneity : 65.0
Completeness : 63.0
V-Measure : 64.0
Silhouette : 0.15650876153333035

Dados teste
---
Homogeneity : 79.0
Completeness : 79.0
V-Measure : 79.0
Silhouette : 0.16472088296509504

Frases por cluster:

<IPython.core.display.HTML object>

---
Documentos por cluster:

Out[59]: <IPython.core.display.HTML object>