projeto

November 21, 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.

```
Using TensorFlow backend.
```

1.1.1 Carregamento dos datasets

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

Contém 33 categorias e ao todo 696 documentos ou frases.

```
                                                                                      pergunt
13   Como posso registrar um cartão SIM?
70   Fui assaltado, preciso acionar seguro.
250  como solicitar um novo certificado digital para um sistema que criei no me departamento?
212  esta dando erro ao tentar criar meu novo username
234  estou tentanto criar meu username, mas não estou conseguindo
179  Por favor me diga quando eu vou ser elegível para minha próxima atualização do dispositiv
223  Eu não sei porque, mas o meu telefone não funciona em casa.
253  Eu tenho um cartão SIM no meu aparelho e eu gostaria de desbloqueá-lo.
77   obrigado mesmo
261  Posso levar meu telefone comigo se eu mudar para uma nova empresa.
```

Qtde. de documentos por categoria:

	Categoria	Qtde
0	ACESSO_REMOTO	20
1	ACTIVATE_DEVICE	19
2	ACTIVATE_PREPAID_PLAN	21
3	ACTIVATE_ROAMING	20
4	ADD_INSURANCE	18
5	ADD_SERVICE_FEATURES	20
6	CERTIFICADO	18
7	CHANGE_PRICE_PLAN	19
8	CONTA	29
9	COVERAGE_AREA_INQUIRY	19
10	DEACTIVATE_PREPAID_PLAN	18
11	DEACTIVATE_ROAMING	18
12	DEVICE_UPGRADE_ELIGIBILITY	24
13	DUVIDAS	17
14	EMAIL	40
15	INTERNATIONAL_RATE_PLAN_INQUIRY	21
16	LINGUAJAR	24
17	NENHUMA_OPCAO	20
18	NETWORK_COMPLAINTS	23
19	NETWORK_UNLOCK	21
20	NOME_BOT	16
21	OBRIGADO	23
22	PORT_IN	21
23	PRICE_PLAN_INQUIRY	23
24	RECHARGE_SIM	18
25	REMOVE_SERVICE_FEATURES	20
26	RETURN_DEVICE	21
27	ROAMING_INQUIRY	20
28	SENHAS	27
29	SERVICOS	20
30	SWAP_DEVICE	15
31	TROUBLESHOOTING	19
32	WIFI	24

```
Total docs     : 696
Total cluster  : 696
X_train size   : (556,)
X_test  size   : (140,)
```

1.1.2 Dataset tokenization

```
Tokenization...
Qtd documentos treino:  556
```

```
Qtd Intents treino   :  33
Finished...
```

```
Out[7]: [TaggedDocument(words=['quer', 'mud', 'servic', 'um', 'nov', 'disposit', 'mant', 'mesm
         TaggedDocument(words=['qual', 'outr', 'servic', 'bot', 'oferec'], tags=[1]),
         TaggedDocument(words=['um', 'nom'], tags=[2]),
         TaggedDocument(words=['possivel', 'algu', 'acompanh', 'atrav', 'etap', 'ativ', 'dispos
         TaggedDocument(words=['precis', 'faz', 'ativ', 'nov', 'telefon', 'pre-pag', 'remov',
         TaggedDocument(words=['nao', 'consegu', 'inic', 'bluetooth', 'emparelh', 'fon', 'ouv'
         TaggedDocument(words=['nenhum', 'opca', 'quer', 'faz'], tags=[6]),
         TaggedDocument(words=['nao', 'precis', 'mais', 'nad', 'obrig'], tags=[7]),
         TaggedDocument(words=['foi-m', 'promet', 'cobertur', 'total', 'cidad', 'inscrev', 'an
         TaggedDocument(words=['acess', 'remot', 'sistem', 'empres', 'cas'], 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       : 100
Infer Epochs : 12000
```

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

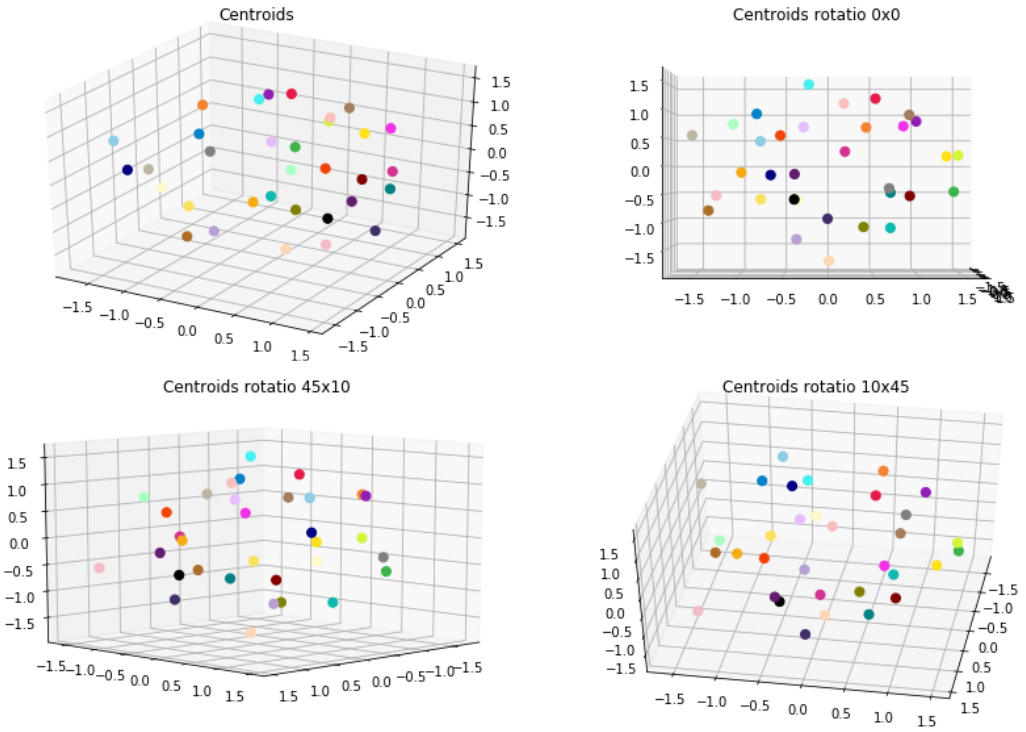
Validação do modelo gerado pelo Doc2Vec... teste tanto nos dados apresentados para treinamento quanto nos dados de testes e as acurácias alcançadas.

```
- Acurácia treino: 98.67
--------------------
 - Acurácia teste 1  : 61.33
 - Acurácia teste 2  : 53.33
 - Acurácia teste 3  : 62.67
 - Acurácia teste 4  : 54.67
 - Acurácia teste 5  : 60.0
- Acurácia média teste: 58.4
```

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:


0     Quanto é o custo de um plano familiar de 3 linhas?
11    O equipamento deve ser devolvido à empresa quando você cancela o meu serviço?
10    Eu gostaria de cancelar o serviço de encaminhamento das minhas ligações. Como faço isso?
8     meu usuario esta cancelado
7     é permitido que patrulheiros tenham uma conta?
..                                                  ...
550   Olá, eu tenho um novo dispositivo, e um cartão SIM de vocês com contrato mensal que obtive
551   Quero sua ajuda para colocar o dinheiro no meu cartão SIM.
552   Como posso registrar um cartão SIM?
553   Eu não posso recarregar porque diz que meu sim está desativado , meu sim pode ser ativado?
555   Quanto é a taxa de ativação de um novo cartão SIM da minha operadora?

[556 rows x 2 columns]


Documentos por cluster:


    C0  C1  C2  C3  C4  C5  C6  C7  C8  C9  C10  C11  C12  C13  C14  C15  C16  C17  C18  C19  C
0   12  17  21  15  20  14  8   12  18  25  12   14   17   25   11   21   26   15   17   19   2
```

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



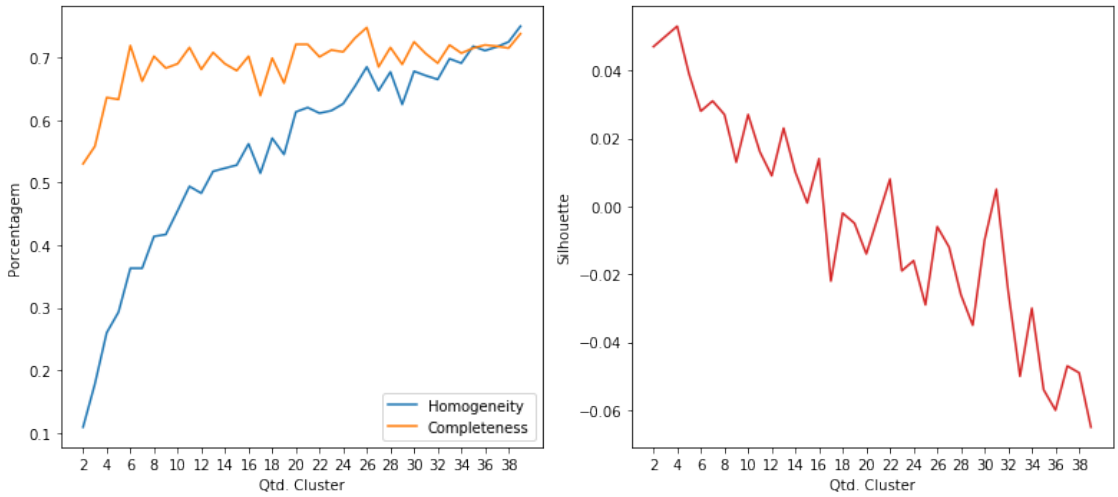
Clusterização dos dados de Teste Por fim, realizada a clusterização dos dados de teste e a apresentação das 8 primeiras frases do conjunto de teste juntamente com outras duas frases do cluster ao qual foi identificado como o melhor.

```
* Quanto da área total no meu país é coberta pela recepção?
  - pqp, seu merda, responde certo porra!
  - Estou planejando uma viagem ao Havaí e estava pensando se eu vou ter sinal lá.
--------------------
* acesso ao meu email foi negado
  - gostaria de informações sobre acesso remoto
  - meu email esta sem acesso
--------------------
* voce nao conseguiu me ajudar, quero falar com um humano
  - Não consigo iniciar o bluetooth para emparelhar com o fone de ouvido, o que devo fazer?
  - voce nao conseguiu me ajudar, quero falar com um atendente
--------------------
* voce nao sabe nada sobre vpn?
  - é para colocar minha senha do meu departamento ou outra?
  - qual usuario e senha usar para acessar os sistemas?
--------------------
* Posso usar o meu telefone quando estou no exterior de férias e quais serão as tarifas?
  - Há alguma limitação quanto aos países no exterior que eu posso chamar e acessar do meu pa
  - Meu telefone pode rastrear quando uso internet no exterior?
--------------------
* Após quanto tempo do pagamento da fatura atrasada, meu chip é desbloqueado?
  - Você pode por favor me informar o status do meu Serviço de Valor Agregado (VAS) - novo to
  - Meu novo telefone não está ativado ainda. Quem pode me ajudar com este problema?
--------------------
* Onde posso localizar informações detalhadas sobre o seu plano de taxa internacional?
  - Poderia por favor compartilhar alguma informação sobre a cobertura de rede na área?
  - Eu vou estar viajando e estou certo se minha empresa fornece cobertura de telefone na min
--------------------
* Posso incluir roaming no meu plano por duas semanas?
  - Quero atualizar o meu plano com o seguro de telefone.
  - Qual é a política de devolução para esse telefone
--------------------
```

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

```
Homogeneidade:  0.709
Completude   :  0.705
Silhouette   :  -0.019
```

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.



1.1.4 TF-IDF

No caso do tf-idf, assim como no doc2vec foi escolhido um máximo de até 1000 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:  556
Qtd Intents treino   :  33
Finished...
```

```
Out[22]: ['quer mud servic um nov disposit mant mesm dar pessoal pod diz faze-l',
          'qual outr servic bot oferec',
          'um nom',
          'possivel algu acompanh atrav etap ativ disposit pre-pag',
          'precis faz ativ nov telefon pre-pag remov restrica',
          'nao consegu inic bluetooth emparelh fon ouv dev faz',
          'nenhum opca quer faz',
          'nao precis mais nad obrig',
          'foi-m promet cobertur total cidad inscrev ano pass mes dois cidad nao conexa conser
          'acess remot sistem empres cas']
```

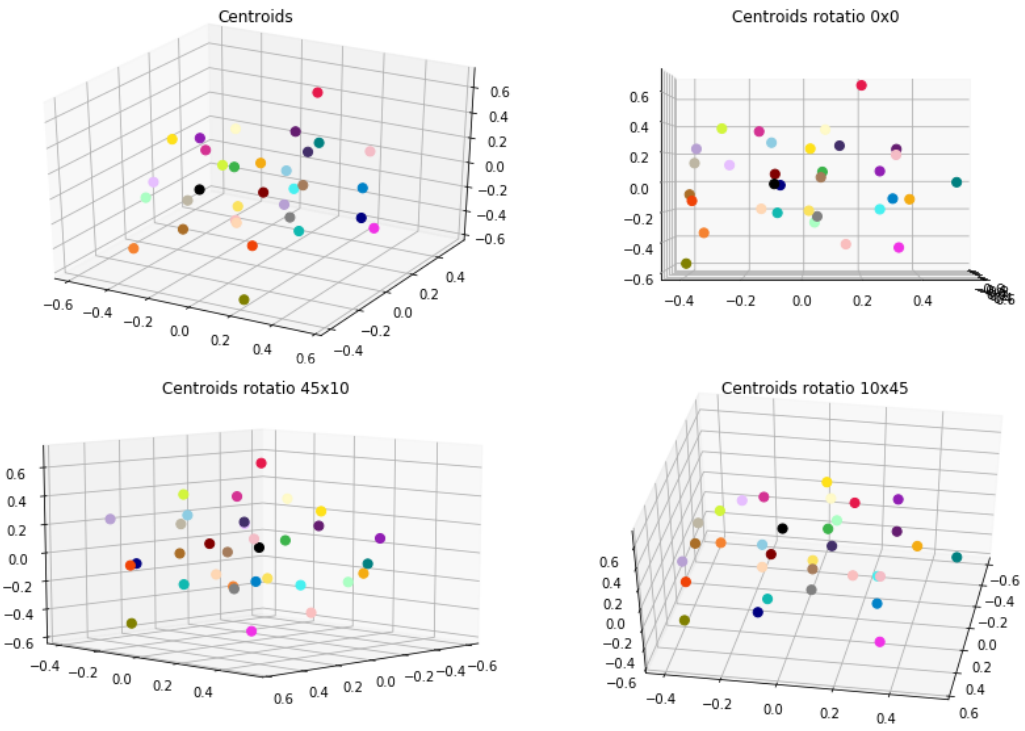
Validação do modelo gerado pelo TF-IDF... teste tanto nos dados apresentados para treinamento quanto nos dados de testes e as acurácias alcançadas.

```
- Acurácia treino: 100.0
--------------------
 - Acurácia teste 1  : 65.33
 - Acurácia teste 2  : 65.33
 - Acurácia teste 3  : 62.67
 - Acurácia teste 4  : 66.67
 - Acurácia teste 5  : 69.33
- Acurácia média teste: 65.87
```

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:



0    você tem um nome?
13   vc tem nome?
12   qual o nome que atribuiram a você?
11   qual é o seu nome?
9    valeu
..     ...
544  Eu estive dirigido muito recentemente e foi praticamente impossível usar o serviço em algu
543  Olá, estava imaginando se essa queda de rede já foi consertada? Já existe uma cobertura L
554  Qual é a área de cobertura e confiabilidade de recepção?
548  Poderia por favor compartilhar alguma informação sobre a cobertura de rede na área?
555  Oi, já que você está aqui, está havendo algum problema com a cobertura na área de SP? Desc

[556 rows x 2 columns]


Documentos por cluster:


   C0  C1  C2  C3  C4  C5  C6  C7  C8  C9  C10  C11  C12  C13  C14  C15  C16  C17  C18  C19  C
0  14  35  21  21  14  10  26  7   14  20  18   11   20   15   19   9    3    23   11   25   1
```

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.

```
* Quanto da área total no meu país é coberta pela recepção?
  - Oi, já que você está aqui, está havendo algum problema com a cobertura na área de SP? Desc
  - Qual é a área de cobertura e confiabilidade de recepção?
--------------------
* acesso ao meu email foi negado
  - estou tentando acessar meu email mas sem sucesso
  - Não consigo entrar no meu e-mail
--------------------
* voce nao conseguiu me ajudar, quero falar com um humano
  - quem e voce?
  - pra que voce serve?
--------------------
* voce nao sabe nada sobre vpn?
  - pra que voce serve?
  - quais outras opcoes voce consegue responder pra mim?
--------------------
* Posso usar o meu telefone quando estou no exterior de férias e quais serão as tarifas?
  - Como salvar contatos do telefone antigo para o telefone novo?
  - posso fazer a configuracao em meu so sem usar a informacao de certificado digital disponil
```

```
--------------------
* Após quanto tempo do pagamento da fatura atrasada, meu chip é desbloqueado?
  - Se eu recarregar R$ 50, o que eu ganho?
  - Qual é o custo por ativação de um dispositivo?
--------------------
* Onde posso localizar informações detalhadas sobre o seu plano de taxa internacional?
  - Gostaria de verificar todas as taxas para cobertura internacional
  - Você pode me dizer a tarifa atual de cobrança da Europa Continental para o Reino Unido do
--------------------
* Posso incluir roaming no meu plano por duas semanas?
  - quero mais informacoes sobre senhas e como proceder
  - Posso incluir mensagens de vídeo como um recurso de serviço novo no meu telefone?
--------------------
```

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

```
Homogeneidade:  0.712
Completude   :  0.732
Silhouette   :  0.043
```