Construindo um chatbot

com Retrieval-based models

Bárbara Barbosa
@bahbbc

Quem sou eu

- Mestra em Sistemas de Informação - com foco em Inteligência Artificial e NLP
- Organizadora do Rails Girls SP, Women Dev Summit e Women in Data Science SP 2019
- Data Scientist Team Leader na Creditas


creditas
vagas.creditas.com.br

Tipos de chatbot

Retrieval vs Generative

Retrieval-based models

Dificuldade: ★★★★★

Vantagens:

+10 de acurácia com o domínio fechado

Pode ser implementado como um conjunto de regras

Desvantagens:

Não lidam bem com textos nunca vistos.

-200 de acurácia com domínio aberto

Generative models

Dificuldade: ★★★★★

Vantagens:

+10 de acurácia com o domínio fechado

Pode parecer que você está falando com uma pessoa

Desvantagens:

Pode falar coisas sem sentido e requer MUITO dado

-50 de acurácia com domínio aberto

Conversas Longas vs Curtas

Conversas longas vs curtas

Conversas curtas

+10 de facilidade

***** Uma resposta por pergunta

Conversas longas

+10 de dificuldade

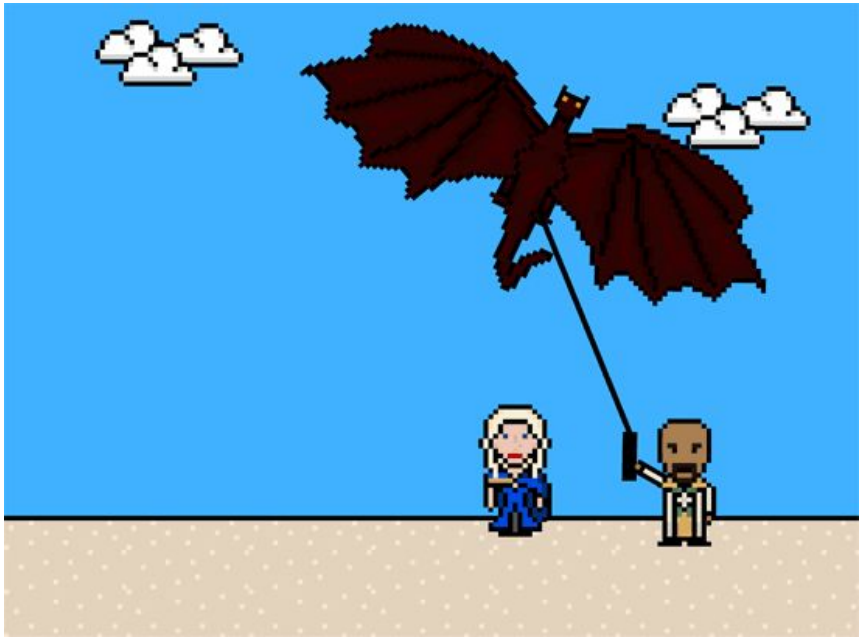
***** É necessário manter histórico do que foi falado.

Missão: Fazer um chatbot usando retrieval models

Obtenha um córpus

Córpus

Obtenha um conjunto de perguntas e respostas.



Córpus

Atenção!

E se eu for criar um córpus?

Córpus

Atenção!

Uma resposta

- pergunta 1
- pergunta 2
- pergunta 3

Você pode usar **aprendizado não supervisionado** para resolver esse problema!

Pré processamento textual

Pré processamento textual

- stemming
- remoção de *stop words*,
- remoção de números,
- caracteres especiais, etc.

Pré processamento textual

Faça as marcações de início e fim de sentença no seu córpus.*

*Não é obrigatório se você não quiser manter o **contexto** da conversa.

Formatação do córpus

Formatação

O seu córpus será diferente para *treino* e *teste*.

No treino você vai ter respostas **verdadeiras** ou **falsas**.

No teste você terá um número *k* de respostas possíveis.

Formatação - Treino

Context	Utterance	Label
blablablabla	bliblibli	1

Dataset de treino

- Como criar respostas falsas?
  - Use respostas que você já possui para fazer as respostas falsas

Dataset de treino

- Como saber quantas criar?
  - Use a proporção original de respostas verdadeiras para escolher a falsa

Dataset de treino

	context	label	utterance
0	nao recib assin trabalh cas cuid bb	0	emprestim garant automovel consegu solicit ate...
1	fac emprestim vcs	1	credit plataform onlin credit missa reduz jur ...
2	qui fac	1	credit plataform onlin credit missa reduz jur ...
3	sim mae	1	credit plataform onlin credit missa reduz jur ...
4	quer sab carr	1	voc precis ajud preenchiment informaco sim pod...

Formatação - Teste

Context	True_utterance	Distractor_1	Distractor_2	Distractor_3
blablablabla	bliblibli	blobloblobo	blublublbu	blebleble

Formatação - Teste

	context	true_utterance	0	1	2
0	imovel valor mil	credit plataform onlin credit missa reduz jur ...	voc precis ajud preenchiment informaco sim pod...	emprestim garant automovel consegu solicit ate...	aqu credit trabalh apen modal emprestim garant...
1	cors classic qt fic financ mil x	emprestim garant automovel consegu solicit ate...	aqu credit trabalh apen modal emprestim garant...	credit plataform onlin credit missa reduz jur ...	voc precis ajud preenchiment informaco sim pod...
2	enta outr empres confavel poss emprest valor	aqu credit trabalh apen modal emprestim garant...	credit plataform onlin credit missa reduz jur ...	emprestim garant automovel consegu solicit ate...	voc precis ajud preenchiment informaco sim pod...
3	exist algum restrica ano veicul	emprestim garant automovel consegu solicit ate...	voc precis ajud preenchiment informaco sim pod...	aqu credit trabalh apen modal emprestim garant...	credit plataform onlin credit missa reduz jur ...
4	voc faz dez mil	credit plataform onlin credit missa reduz jur ...	aqu credit trabalh apen modal emprestim garant...	voc precis ajud preenchiment informaco sim pod...	emprestim garant automovel consegu solicit ate...

Treinamento do modelo

Construindo um classificador

- Classificador binário

- O classificador roda *n* vezes, sendo *n* o número possível de respostas.

Construindo um classificador

- Escolher uma técnica de representação de palavras
  - Embeddings (word2vec)
  - Bag of words (n-gramas)

Construindo um classificador

- Classificador multiclasse

- Neste caso não haveria necessidade de certos pré-processamentos

Construindo um classificador

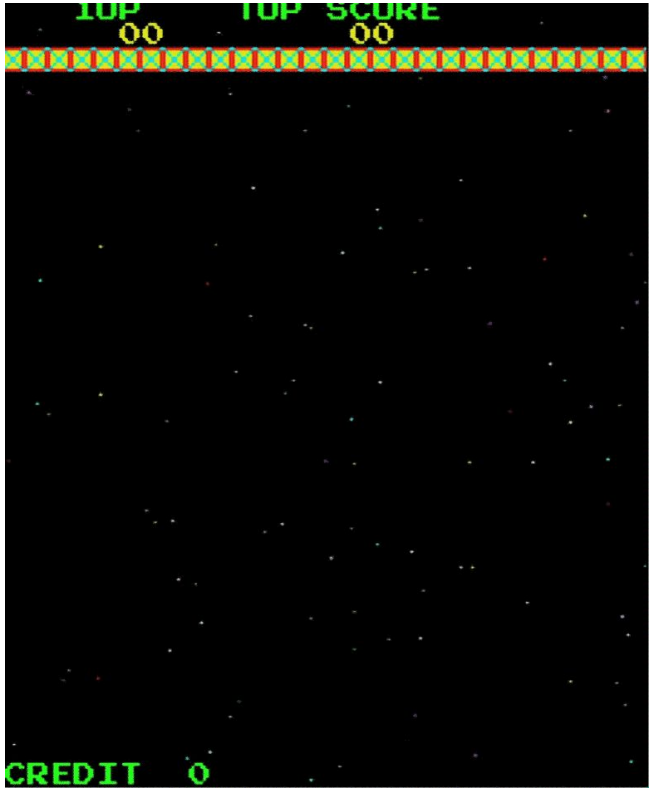
Você pode:

- Utilizar a mesma estrutura do paper do UBUNTU DIALOG CORPUS e fazer com LSTMs.
- Utilizar uma outra estrutura, como *Random Forests*.

Referências

- **Ubuntu Dialogue Corpus** https://arxiv.org/pdf/1506.08909.pdf
- **Chatbot tutorial** http://bit.ly/2FRur9f
- **Tutorial de filmes com word2vec (RF)** https://github.com/bahbbc/movie-plots-by-genre

Obrigada!



Dúvidas?

Bárbara Barbosa

Twitter: @bahbbc

Vagas: vagas.creditas.com.br

LinkedIn: br.linkedin.com/in/bahbbc

Slides:
bit.ly/tdc-chatbot-ia