THE DEVELOPER'S CONFERENCE

Quer ter um bom chatbot?
Então vamos começar entendendo um pouco de matemática!

Renato Barbosa
RnD Innovation Architect

Fernando Sapata
Enterprise Solutions Architect

Globalcode

Agenda

THE DEVELOPER'S CONFERENCE

- ¿
- Falando a mesma língua
- NLUs e seus algoritmos
- Métricas
- Quero mudar de engine...
- Testando
- Conclusão

Globalcode

Falando a mesma língua

Vocabulário do Chatbot

THE DEVELOPER'S CONFERENCE

Utterances

- Qualquer coisa que o usuário diga.
- Ex: "*Quero as notícias de ontem*"

Intent

- Intenção do usuário
- *Obter notícias*

Globalcode

Vocabulário do Chatbot

THE DEVELOPER'S CONFERENCE

Entidade / Slot / Parâmetro:

- Qualifica / especifica uma intenção, são:
  - Nomes
  - Data
  - Tipos de informação

- Ex: "*Quero as notícias de **ontem** sobre o **TDC***"

Globalcode

Stop Words

THE DEVELOPER'S CONFERENCE

- Palavras com pouco significado, como:
  - "e", "o", "a", "uma", "que", "com"
- Palavras comuns em um idioma.

Globalcode

Tf-Idf

THE DEVELOPER'S CONFERENCE

- Método Estatístico
- Term Frequency–Inverse Document Frequency
- Identifica a relevância de uma palavra no documento

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Globalcode

Tokenização e Tokens

THE DEVELOPER'S CONFERENCE

- Tarefa de dividir uma sentença
- Ex: “Olá TDC São Paulo.”

  Tokens: [“Olá”, “**TDC**”, “São”, “Paulo”]

  Array de palavras, não respeita ordem.

Globalcode

N-Grams

THE DEVELOPER'S CONFERENCE

- Math: Método probabilístico que usa sequência de **N** palavras tokenizadas.
- Tipos, "Olá TDC São Paulo.":
  - Unigram: "Olá", "**TDC**", "São", "Paulo"
  - Bi-gram: "Olá TDC", "TDC São", "**São Paulo**"
  - Tri-gram: "Olá TDC São" "**TDC São Paulo**"
  - Quad-gram "Olá TDC São Paulo"

Globalcode

Modelos Enviesados

THE DEVELOPER'S CONFERENCE

- Math: Modelo estatístico no qual classes não possuem equilíbrio.
- Ex: Bot com 3 intenções
  - Intenção 1 – 10 Uterrances
  - Intenção 2 – 5 Uterrances
  - Intenção 3 – 45 Uterrances
- Que diferença isso faz?

Globalcode

Distâncias

THE DEVELOPER'S CONFERENCE

- Algoritmos utilizados para representar o quão parecidas são as sentenças / strings.
- Levenshtein
  - Considera o número de edições por substituição
- Hamming
  - Considera número de posições nas quais as strings diferem entre si

Globalcode

NLUs e seus algoritmos

Algoritmos de Classificação

THE DEVELOPER'S CONFERENCE

- Word2Vec
- Bag of Words
- FST
- Bayes
- SVM
- CNN
- Dynamic Word Weight

Globalcode

Métricas

Acurácia

THE DEVELOPER'S CONFERENCE

- É a métrica mais simples de ser calculada
- Basta dividir o total de acertos pelo total de amostras

Ex: 150 (acertos) / 186 (amostras) = 80,64%

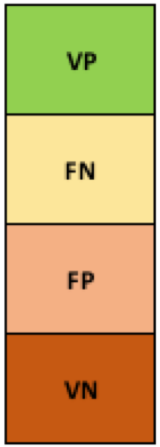
Parece bom!

Globalcode

Matriz de Confusão

THE DEVELOPER'S CONFERENCE

- É uma tabela que permite visualizar o desempenho do classificador.



	Correta		
Predita	Sim	Não	Total
Sim	60	32	92
Não	76	18	94
Total	136	50	186

Globalcode

Precision

THE DEVELOPER'S CONFERENCE

- O objetivo da métrica é identificar quantas amostras foram classificadas positivamente.

$$P = \frac{VP}{VP+FP}$$

P = 60 / (60 + 32)
P = 0.65

Globalcode

Recall

THE DEVELOPER'S CONFERENCE

- Conhecido como taxa de sensibilidade.
- Tem a mesma ideia da *precision*, porém para as amostras **falsas negativas**.

$$R = \frac{VP}{VP+FN}$$

R = 60 / (60 + 76)
R = 0.44

Globalcode

F1

THE DEVELOPER'S CONFERENCE

- O *F1 score* é definido como duas vezes a média harmônica entre *R e P*, ou seja, é um 'meio termo' entre as duas métricas anteriores

$$F1 = 2 \times \frac{P \times R}{P + R}$$

F1 = 2 x ((0.65 x 0.44) / (0.65 + 0.44))
F1 = 2 x (0.286 / 1.09)
F1 = 2 x 0.26
F1 = 0.52

Globalcode

+ Métricas

THE DEVELOPER'S CONFERENCE

- % entidades identificadas
- Performance da classificação (velocidade)
- Performance na resolução das entidades
- Performance de chamadas externas
- % de confiança da classificação

Globalcode

Testando

ATIS Dataset

THE DEVELOPER'S CONFERENCE

Airline Travel Information System

- Foi coletado pela DARPA[1] no início dos anos 90
- Possui 4978 sentenças de treinamento e;
- 893 sentenças de teste
- 25 Classes
- Diversos slot types diferentes

http://lisaweb.iro.umontreal.ca/transfert/lisa/users/mesnilgr/atis/

1 Defense Advanced Research Projects Agency (DARPA)

Globalcode

Dataset Puro

THE DEVELOPER'S CONFERENCE

- ATIS usa representação IOB
- BOS e EOS (Begin / End of Sentence)
- B-I (Begin of Entity - Increment of Entity)

- Ex: (list the **delta airlines** flights from **boston** to **Philadelphia**)
- BOS list the **delta airlines** flights from **boston** to **philadelphia** EOS
- O O O **B-airline_name I-airline_name** O O **B-fromloc.city_name** O **B-toloc.city_name** O

Globalcode

Preparando o dataset

THE DEVELOPER'S CONFERENCE

- Limpar utterrances inválidas (limites de NLU)
- Remover Uterrances duplicadas
- Converter :
  - Classes em intenções (class mapping)
  - Entidades identificadas em entidades do modelo (entity mapping)
  - Criar lista de utterances do modelo com entity map
  - Criar lista de utterances de testes sem entity map
- Gerar JSON para treinar a sua NLU

Globalcode

Avaliando o score

THE DEVELOPER'S CONFERENCE

- Cuidado: Sem nenhum esforço você já acerta 71%
- if (F1 > ~97%) => alert('Será!?');
- Repita os testes em engines diferentes
- Repita os testes com estratégias diferentes

Globalcode

Quero mudar de engine...

goto: slide 1

Conclusão

THE DEVELOPER'S CONFERENCE