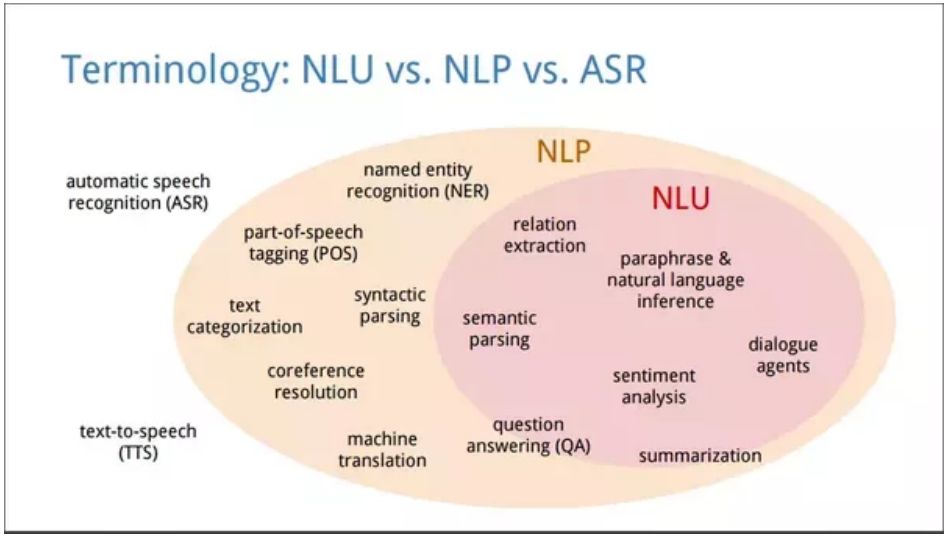
Natural Language Processing and Data Science: iFood Applications

Francielle Vargas

June 9, 2018

Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

- Intersection between Computer Science and Linguistics;
- It's not Machine Learning;



Purpose

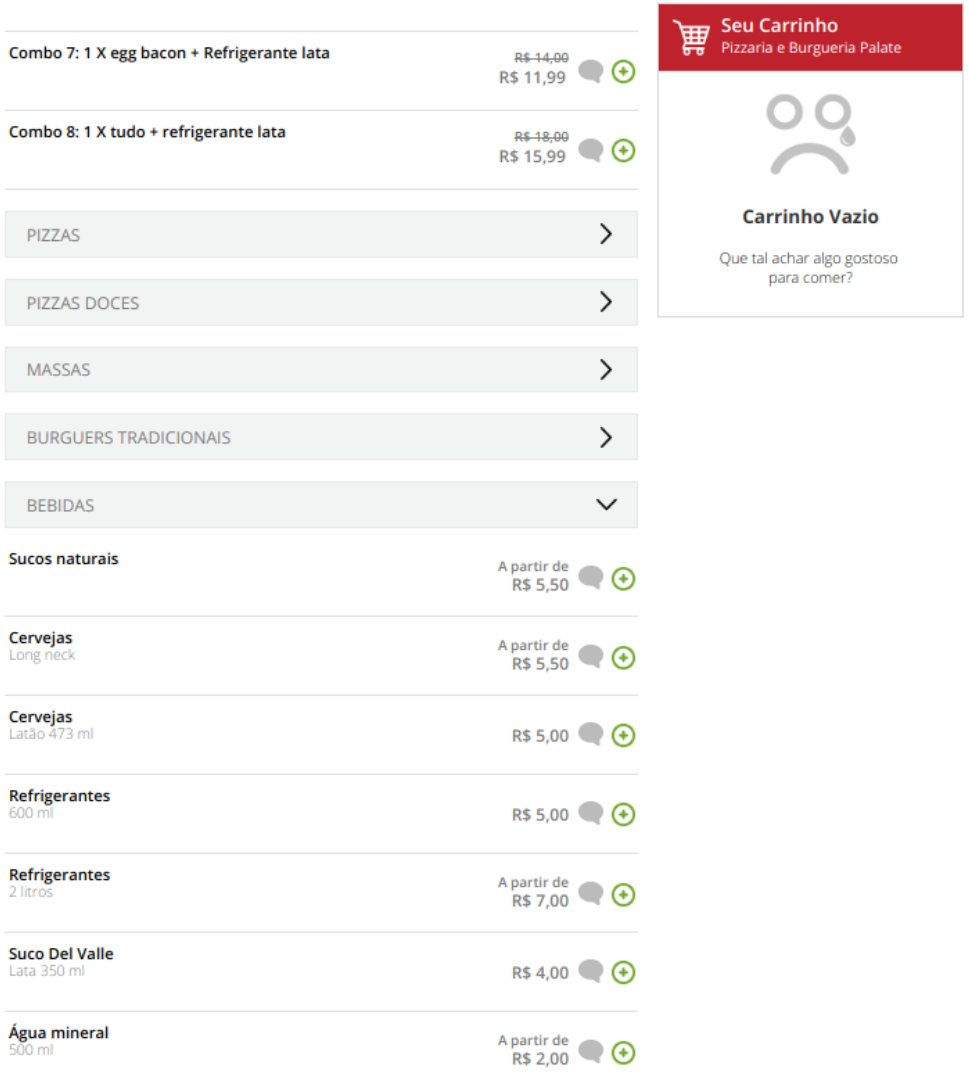
- Overview about NLP applied to Data Science;
- Challenges Data Science (NLP) in the iFood;
- Preparing / Presentation of the development environment (python);
- Run one or two classic methods.

Natural Language Processing

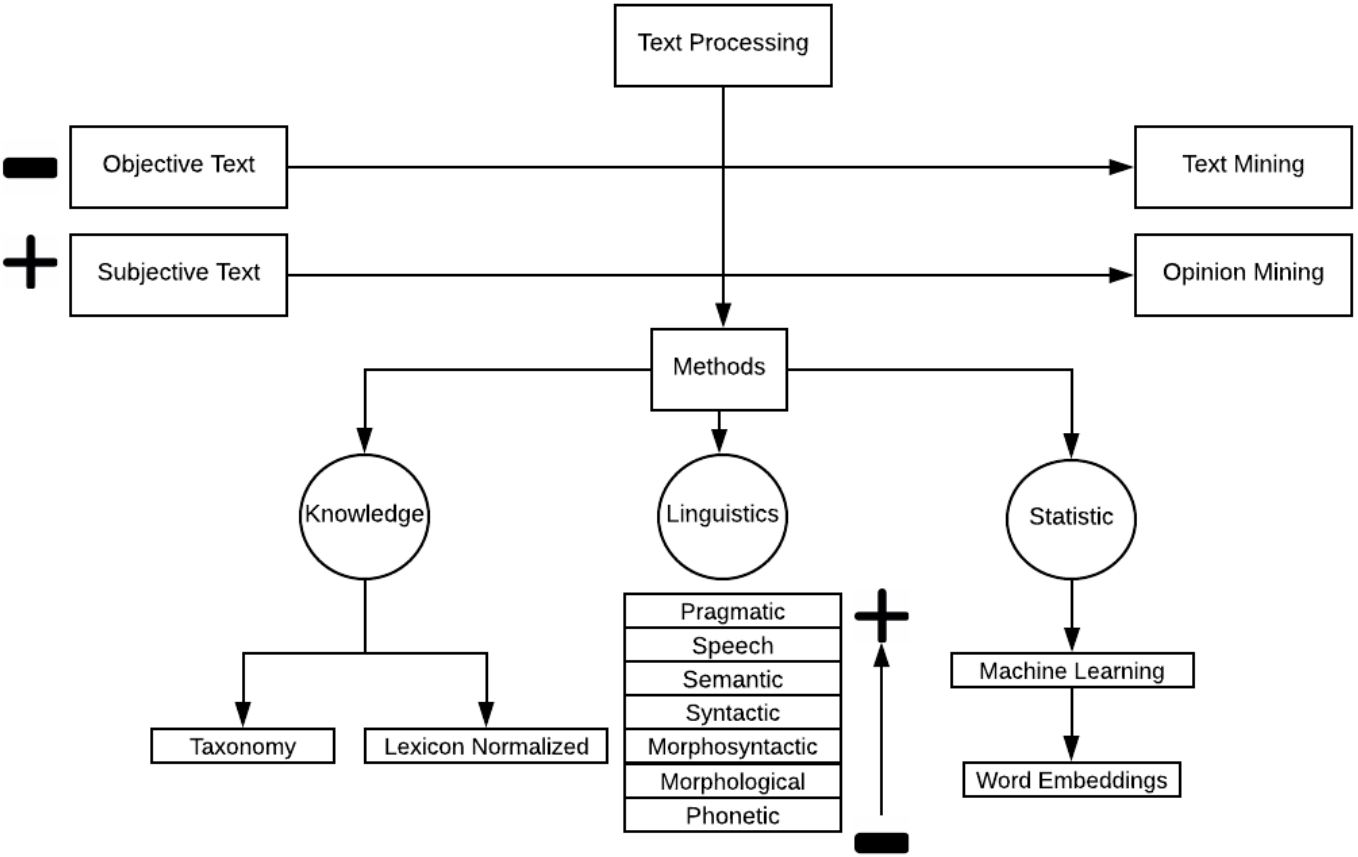
- Word categorization and tagging;
- Syntactic parsing;
- Topic modeling;
- Application of machine learning;
- Semantic similarity and clustering;
- Short phrases/notes semantic analysis;
- Text matching and similarity;
- Word embedding;
- Lexicon normalization;
- Named entity recognition;
- Summarization;
- Sentiment Analysis or Opinion Mining.

Problem


Pizzaria e Burgueria Palate
4,6
Comida Brasileira • $$
5276.38 km • 45 min
Pizzaria e Burgueria Palate
Informações Extras
Cardápio
Avaliações
Filtrar Menu
Prato, ingrediente, etc...
Seu Carrinho
Pizzaria e Burgueria Palate
PROMOÇÃO
2 pizza gigante 10 fatias 35 cm + guarana 2 litros
R$ 92,96
R$ 61,99
pizza grande 8 fatias ,30cm + guarana 1 litro
A partir de
R$ 24,90
2 pizza gigante 10 fatias 35 cm + guarana 2 litros
A partir de
R$ 123,78
2 pizzas grandes 8 fatias 30 cm + guarana 2 litros
R$ 54,80
R$ 49,99
Pesunto gigante
Pizza gigante 35 cm 8 fatias
R$ 30,90
R$ 26,99
Carrinho Vazio
Que tal achar algo gostoso para comer?


Combo 7: 1 X egg bacon + Refrigerante lata
R$ 14,00
R$ 11,99
Combo 8: 1 X tudo + refrigerante lata
R$ 18,00
R$ 15,99
PIZZAS
PIZZAS DOCES
MASSAS
BURGUERS TRADICIONAIS
BEBIDAS
Sucos naturais
A partir de
R$ 5,50
Cervejas
Long neck
A partir de
R$ 5,50
Cervejas
Latão 473 ml
R$ 5,00
Refrigerantes
600 ml
R$ 5,00
Refrigerantes
2 litros
A partir de
R$ 7,00
Suco Del Valle
Lata 350 ml
R$ 4,00
Água mineral
500 ml
A partir de
R$ 2,00
Seu Carrinho
Pizzaria e Burgueria Palate
Carrinho Vazio
Que tal achar algo gostoso para comer?

Overview


Text Processing
Objective Text
Subjective Text
Text Mining
Opinion Mining
Methods
Knowledge
Linguistics
Statistic
Taxonomy
Lexicon Normalized
Pragmatic
Speech
Semantic
Syntactic
Morphosyntactic
Morphological
Phonetic
Machine Learning
Word Embeddings

OpCluster-PT: Automatic taxonomy creating

Algoritmo 7: Algoritmo OpCluster-PT

Entrada: Lista de aspectos $A = \{a_1, a_2, ..., a_n\}$ ordenados de forma decrescente por critério de frequência;
Revisões processadas pelo CORP $R = \{r_1, r_2, ...r_n\}$, em que os aspectos de A ocorrem;
Saída: Grupos de aspectos $G = \{g_1, g_2, ..., g_n\}$, tal que cada g_i contém subconjuntos de aspectos de A;

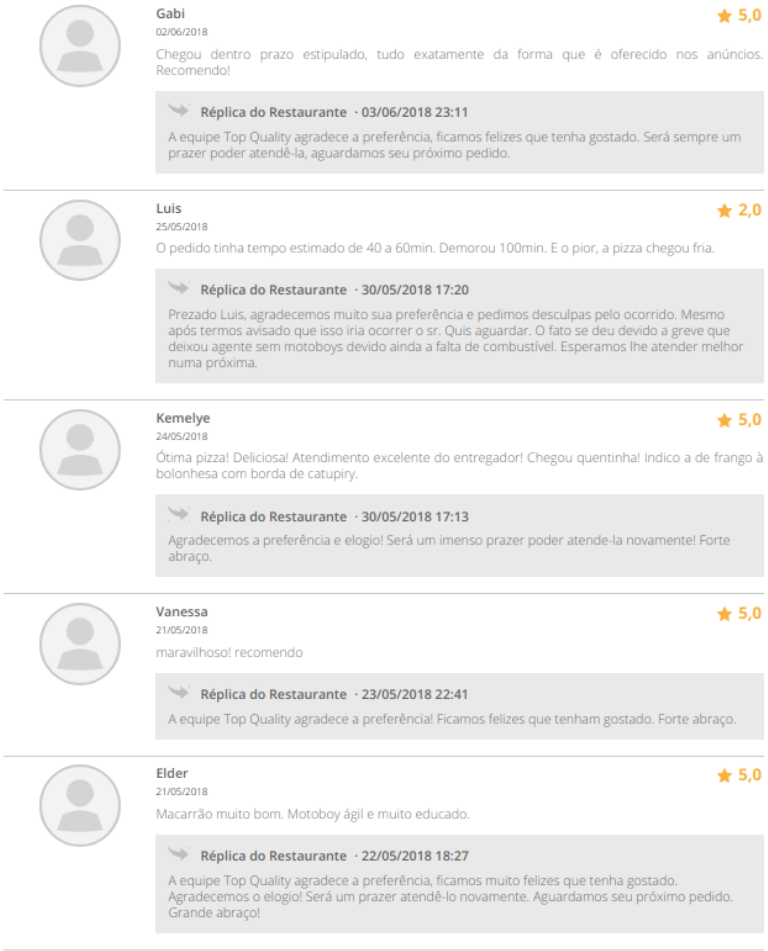
início
  Declare $\mathbf{B} = \{b_{sin}, b_{parte\text{-}todo}, b_{causa}, b_{deverb}, b_{estrag}, b_{diminu}, b_{corref}, b_{subst}\}$, tal que **B** contém o resultado da busca por aspectos em relação de sinonimia, meronímia/holonímia, causativa e construções deverbais, correferentes, estrangeirismos, diminutivos (por exemplo, b_{sin} contém os aspectos sinônimos ao aspecto de interesse);
  Declare $\mathbf{U} = \{u_1\ u_2, ...\ u_n\}$, tal que cada conjunto u_i contém um grupo unitário de **G**;
  Declare *contador* = 0;
  Declare *posicao* = 0;
  repita
    se a_i *de A possuir sinônimos na base do Onto.PT* **então**
      Adiciona em b_{sin} os sinônimos encontrados;
    fim
    se a_i *de A possuir merônimos e/ou holônimos imediatos na base do Onto.PT* **então**
      Adiciona em $b_{parte\text{-}todo}$ os merônimos e/ou holônimos encontrados;
    fim
    se a_i *de A possuir relações causativas do tipo resultadoDaAção e/ou serveParaAccao na base do Onto.PT* **então**
      Adiciona em b_{causa} os itens em relação *resultadoDaAção e/ou serveParaAccao* encontrados;
    fim
    se a_i *de A possuir construções deverbais na base do iLteC* **então**
      Adiciona em b_{deverb} as construções deverbais encontradas;
    fim
    se a_i *de A possuir estrangeirismos na base do iLtеC* **então**
      Adiciona em b_{estrag} os estrangeirismos encontrados;
    fim
    se a_i *de A possuir construções de diminutivos na lista de diminutivos/aumentativos* **então**
      Adiciona em b_{diminu} os diminutivos encontrados;
    fim
    se a_i *de A possuir relações de substring com outros aspectos de A* **então**
      Adiciona em b_{subst} os aspectos em relações de *substring* encontradas;
    fim
    se a_i *de A, nas revisões em que ocorre, possuir correferentes classificados pelo CORP* **então**
      Adiciona em b_{corref} as cadeias de correferentes encontradas;
    fim
    Exclua itens duplicados de $\mathbf{B} = \{b_{sin}, b_{parte\text{-}todo}, b_{causa}, b_{deverb}, b_{estrag}, b_{diminu}, b_{corref}, b_{subst}\}$, se houver;
    Incremente *contador*;
    Crie grupo G_i e adicione em G_i os aspectos da intersecção (A,B);
    Remova de A os aspectos da intersecção;
    Esvazie **B**;
    repita
      se *aspecto de G, nas revisões em que ocorre, possuir correferentes classificados pela aplicação CORP* **então**
        Adiciona em b_{corref} as cadeias de correferentes encontradas;
      fim
      se *aspecto de G possuir estrangeirismos na base do iLtеC* **então**
        Adiciona em b_{estrag} os estrangeirismos encontrados;
      fim
      se *aspecto de G possuir construções de diminutivos na lista de diminutivos/aumentativos* **então**
        Adiciona em b_{diminu} os diminutivos encontrados;
      fim
      Exclua itens duplicados de $\mathbf{B} = \{b_{corref}, b_{estrag}, b_{diminu}\}$, se houver;
      Adicione em G_i os aspectos da intersecção (A,B);
      Remova de A os aspectos da intersecção;
      Esvazie **B**;
      Guarde em **posição** a última posição do elemento adicionado em G_i
    até *a posição dos elementos de G for maior que valor de posição*;
  até *A esvaziar*;
  repita
    Selecione os grupos unitários e adicione em U_i;
    se U_i *estiver contido em aspectos de* G_i *por relação de substring* **então**
      Adicione em G_i o aspecto de U_i;
      Remove U_i de G
    fim
  até *G esvaziar*;;
fim

Some Methods

Methods

- Step 1: Tokenization;
- Step 2: Lemmatization;
- Step 3: Taxonomy learning;
- Step 4: Named Entity Recognition;
- Step 5: Labeled menu data.

Opinion Mining Problem



Gabi
02/06/2018
★ 5,0
Chegou dentro prazo estipulado, tudo exatamente da forma que é oferecido nos anúncios. Recomendo!

Réplica do Restaurante · 03/06/2018 23:11
A equipe Top Quality agradece a preferência, ficamos felizes que tenha gostado. Será sempre um prazer poder atendê-la, aguardamos seu próximo pedido.

Luis
25/05/2018
★ 2,0
O pedido tinha tempo estimado de 40 a 60min. Demorou 100min. E o pior, a pizza chegou fria.

Réplica do Restaurante · 30/05/2018 17:20
Prezado Luis, agradecemos muito sua preferência e pedimos desculpas pelo ocorrido. Mesmo após termos avisado que isso iria ocorrer o sr. Quis aguardar. O fato se deu devido a greve que deixou agente sem motoboys devido ainda a falta de combustível. Esperamos lhe atender melhor numa próxima.

Kemelye
24/05/2018
★ 5,0
Ótima pizza! Deliciosa! Atendimento excelente do entregador! Chegou quentinha! Indico a de frango à bolonhesa com borda de catupiry.

Réplica do Restaurante · 30/05/2018 17:13
Agradecemos a preferência e elogio! Será um imenso prazer poder atende-la novamente! Forte abraço.

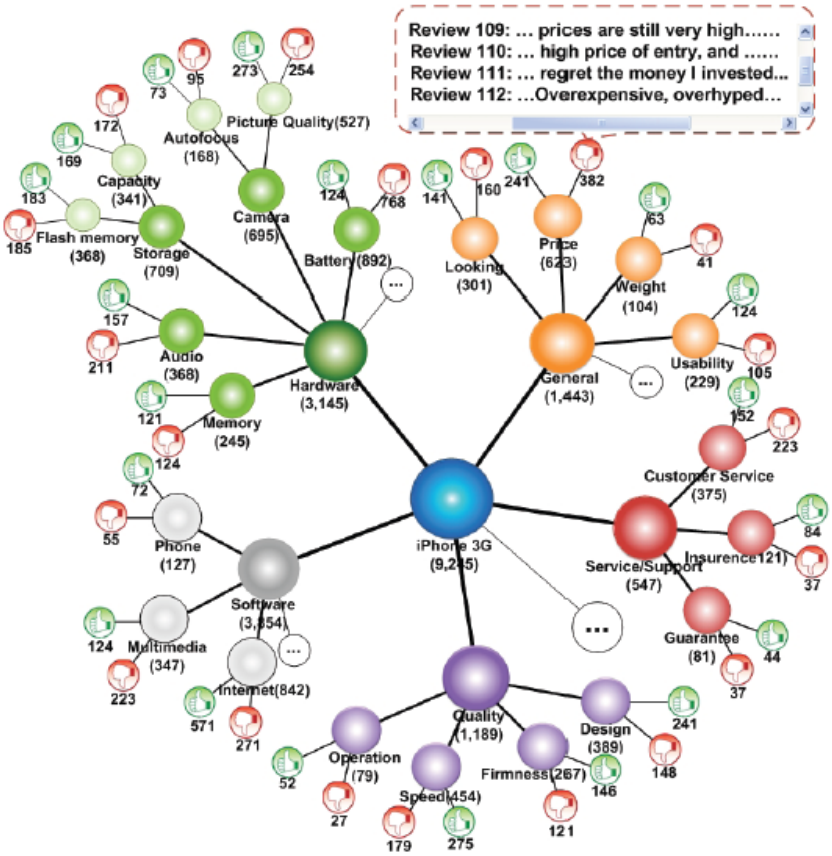
Vanessa
21/05/2018
★ 5,0
maravilhoso! recomendo

Réplica do Restaurante · 23/05/2018 22:41
A equipe Top Quality agradece a preferência! Ficamos felizes que tenham gostado. Forte abraço.

Elder
21/05/2018
★ 5,0
Macarrão muito bom. Motoboy ágil e muito educado.

Réplica do Restaurante · 22/05/2018 18:27
A equipe Top Quality agradece a preferência, ficamos muito felizes que tenha gostado. Agradecemos o elogio! Será um prazer atendê-lo novamente. Aguardamos seu próximo pedido. Grande abraço!

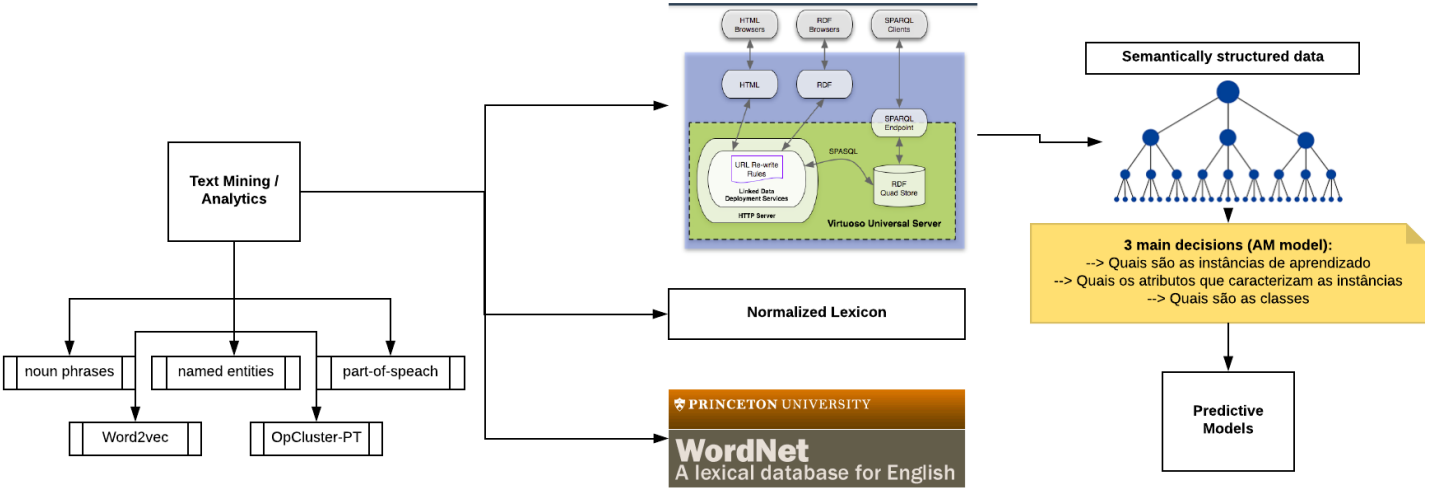
Opinion Mining Problem



Papers - Similar Approach

- WordNet — A Lexical Database for English, `http://wordnet.princeton.edu/`
- Rychtyekyj, N. *DLMS: An Evaluation of KL-ONE in the Automobile Industry.* **Ford Motor Company, Manufacturing Quality Business Systems**. AAAI Technical Report WS.
- Aciar, S.; Zhang, D.; Simoff, S. and Debenham, J.*Recommender System Based on Consumer Product Reviews.*Proceedings of the International Conference on Web Intelligence, 2006.
- Faure, D. and Nédellec, C. *A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition.* Laboratoire d'Intelligence Artificielle de Paris V, LREC 2018.

Final Architecture



First steps

- Create GitHub Account (https://github.com/);
- Install Linux or install Python;
- Install a IDE (Sublime);
- Start coding.

Thank you very much ;)

franciellealvargas@gmail.com