Extração de conceitos de textos

Trabalho de IA

Diego Rubin – http://diegorubin.com

Motivação

- Maior banco de dados existente;
- Algo realmente útil;
- Caso ocorrido na Exame;
- Desafio.

Objetivo do Trabalho

- Encontrar termos conhecidos no texto;
- Exibir resultados gerados a partir de processos realizados;
- Classificação automática.

Etapas do Processo

- Separação em tokens
- Remoção de palavras inúteis
- Radicalização das palavras
- Part-of-speech Tagger
- Busca em um dicionário

Separação em tokens

- Delimitadores
  - Espaços em branco
  - Sinais de pontuação
- Obtenção de um vetor de palavras
- Exemplo:
  - “Python é uma linguagem legal.”
  - [”Python”, ”é”, ”uma”, ”linguagem”, ”legal”, ”.”]

Remoção de palavras inúteis

- Palavras sem valor semântico
- Também conhecidas como stopwords
- Lista pré-definida
- Primeira versão provida pela NLTK
- Exemplos
  - Mais
  - Por
  - Como

Radicalização das palavras

- Diminuição na variação de palavras
- Algoritmo chamado RSLP(Removedor de Sufixos da Língua Portuguesa)
- Desenvolvido por Vivana Orengo
- Utiliza um dicionário de exceções.

RSLP – 1ª Etapa

- Redução do plural
- Remove-se o final -s indicativo de plural de palavras que não se constituem em exceções a regra, realizando modificações, quando necessário
- Algumas exceções:
  - Lápis
  - Férias

RSLP – 2ª Etapa

- Redução do feminino
- Remove-se o final -a de palavras femininas com base nos sufixos mais comuns.

RSLP – 3ª Etapa

- Redução adverbial
- Remove-se o final -mente de palavras que não se constituem em exceção.

RSLP – 4ª Etapa

- Redução aumentativo/diminutivo
- Removem-se os indicadores de aumentativo e diminutivo mais comuns.

RSLP – 5ª Etapa

- Redução nominal
- Removem-se 61 sufixos possíveis para substantivos e adjetivos.

RSLP – 6ª Etapa

- Redução verbal
- Reduzem-se as formas verbais aos seus radicais.

RSLP – 7ª Etapa

- Remoção de vogais
- Removem-se as vogais *a*, e *e* o das palavras que não foram tratadas pelos dois passos anteriores.

RSLP – 8ª Etapa

- Remoção de acentos
- Removem-se os sinais diacríticos das palavras.

Part-of-Speech Tagger

- Linguística de Corpus
- Corpus
- Floresta
- Treinamento do Tagger
- Exemplo:
  - “O rato roeu a roupa”
  - “O/a rato/n roeu/v a/a roupa/n”

Freebase

- Base de informações colaborativas
- Utilizado para buscar conceitos
- Excelente API


```
{"status": "200 OK",
  "result": [
      {
          "mid": "/m/0djmwv",
          "id": "/en/arduino",
          "name": "Arduino",
          "notable": {
              "name": "Computing Platform",
              "id": "/computer/computing_platform"
          },
          "lang": "en",
          "score": 50.266792
      }]
}
```

Serviço REST

- Protocolo cliente/servidor sem estado
  - HTTP
- Conjunto de operação bem definidas
  - GET, POST e PUT
- Identificação única para os recursos
  - /documents/id
- Uso de hipermídia para representação dos estados ou transições
  - JSON

Técnologia utilizada no serviço

- Python
  - Facilidade de manipulação de strings
  - Dinâmica
  - Google
- Twisted
  - Servidor HTTP
  - Simples
- NLTK
  - Auxilia nas tarefas

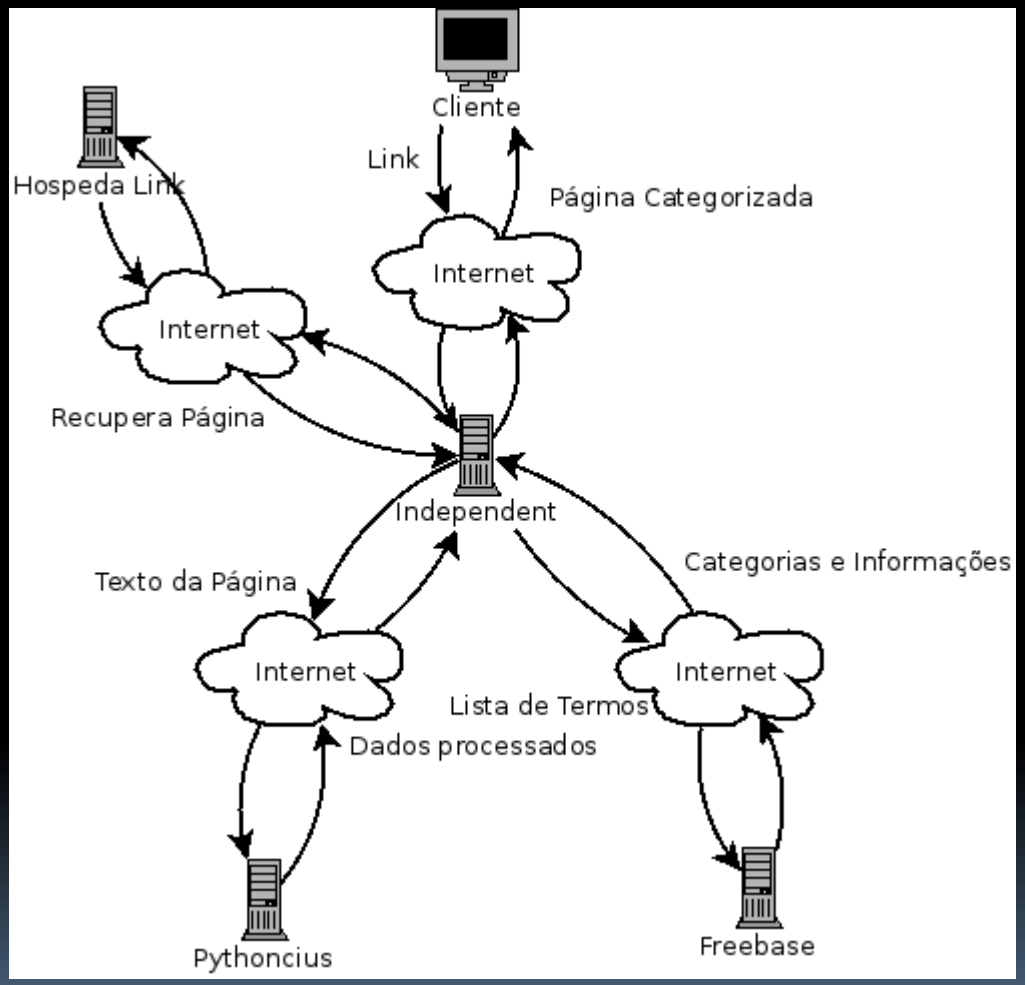
Persistência dos Dados

- Utilização MongoDB
- NoSQL
- Documents e Collections
- JSON

Aplicações Criadas

- Servidor(Pythonicus)
- Cliente(Demo)
- Cliente(Classificador)
  - CMS Independent
  - Baseados nos resultados do Servidor, busca conceitos no Freebase

Classificador



Correções

- Problemas com acentuação;
- Melhorar forma de limpeza das páginas;
- Melhorar treinamento do Corpus.

Conclusões

- Tarefa não trivial;
- Um ponto de partida;
- Bons resultados obtidos no domínio escolhido.