Linguagens de Anotação de Documentos

2016/17

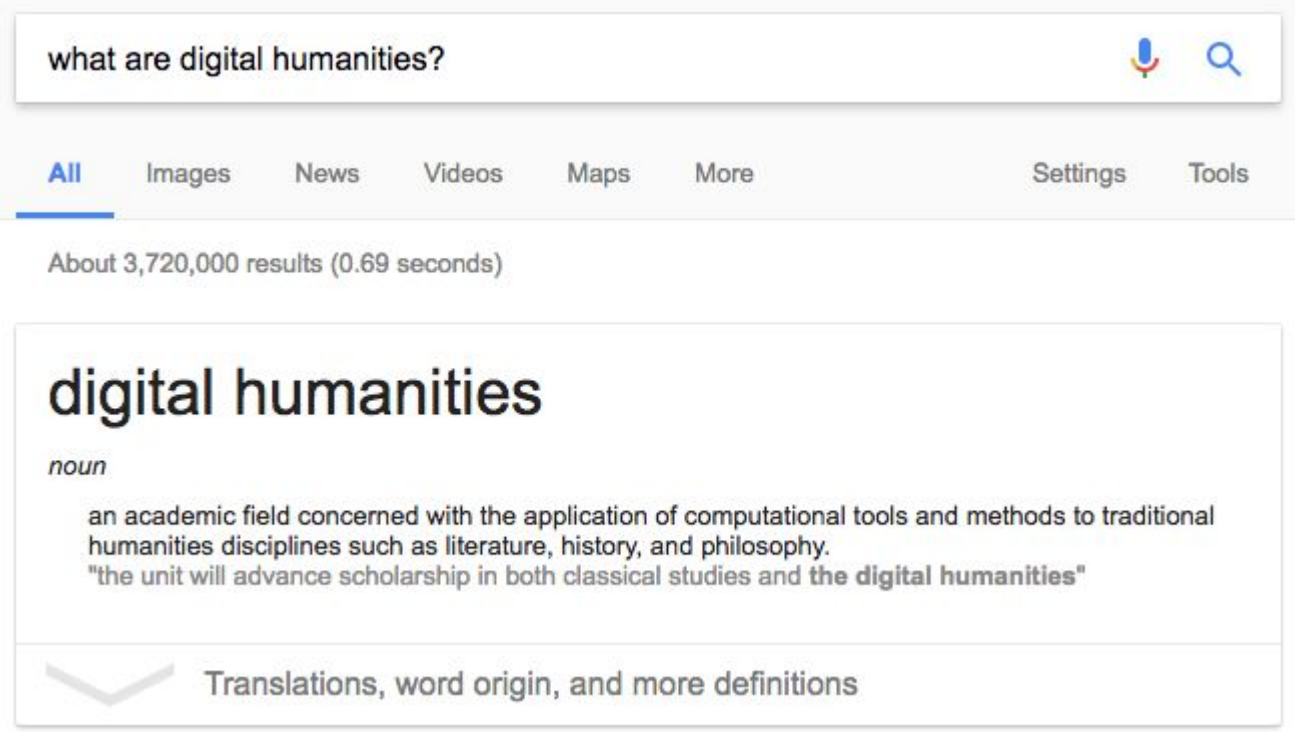
Apresentação

Porque é que isto vos interessa?

Motivação

- A um nível ou outro, nada é feito hoje sem apoio **tecnológico**, e a linguística não é exceção
- A um nível mais baixo, é essencial para que a **partilha** de informação e o trabalho **colaborativo** seja eficaz
- A um nível mais avançado, permite **automatizar** tarefas linguísticas ou extração de **novo** conhecimento
- Isto requer normas e práticas uniformes para representar e manipular **informação**
- Mas a relação é bidirecional: os cientistas da computação dependem hoje em dia da linguística em várias tarefas que lidam com **linguagens naturais**

Humanidades Digitais



Google

Humanidades Digitais

Google

Humanidades Digitais

- Aplicação de ferramentas computacionais no contexto das humanidades
- *Primeira vaga:* tecnologia como **suporte** às humanidades
  - Quantitativa: pesquisa, arquivo, classificação, ...
  - Infraestrutura para representar e partilhar conhecimento
- *Segunda vaga:* tecnologia no **centro** do processo
  - Qualitativa: interpretação, análise, inferência, …
  - Analisa conhecimento atual para criar novo conhecimento
- Sem a **estrutura** imposta pelas técnicas de fundo não seria possível aplicar as técnicas avançadas de análise

Humanidades Digitais

- *Primeira vaga:*
  - Corretores ortográficos, corretores gramaticais
  - Digitalização e preservação de documentos
  - Geração de documentos digitais
  - Visualização estruturada de informação
- *Segunda vaga:*
  - Mineração de texto
  - Traduções assistidas
  - Processamento de linguagem natural
  - Reconhecimento e síntese de voz
- Nesta cadeira vamos focar-nos em técnicas da primeira fase

Linguística Computacional

“...the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena.”

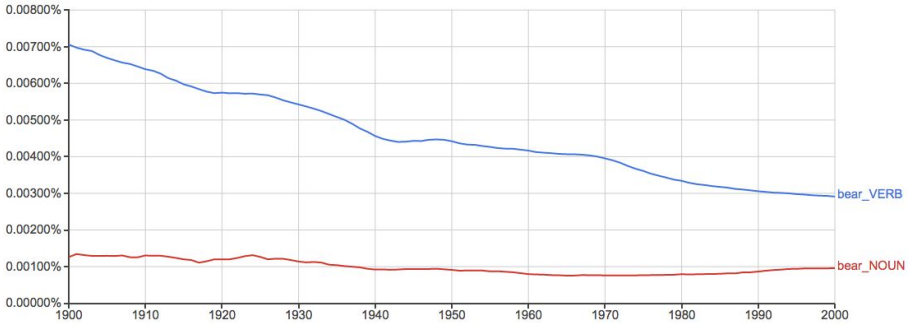
Association for Computational Linguistics

- Contribuições da linguística para as ciências da computação
- Foca-se na análise computacional de informação em linguagem natural para processar e sintetizar **textos** ou **voz**

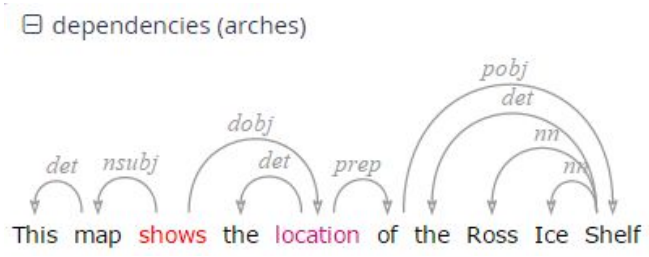
Aplicações Concretas

Exemplo: *Corpora* Anotados

- *Corpora* de linguagens com anotações
- Desafios linguísticos:
  - Estruturação da informação
  - Classificação de partes do texto
  - Técnicas de procura e consulta



Google Ngram Viewer



Annis

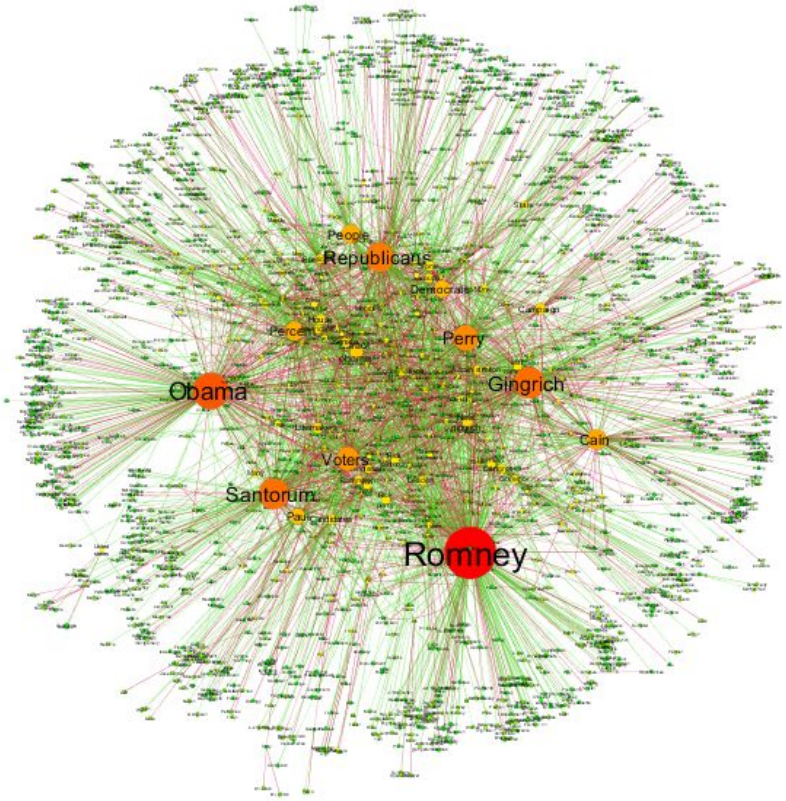
Exemplo: Tradução Automática

- Tradução automática de textos entre linguagens naturais
- Desafios linguísticos
  - Sintaxe
  - Semântica
  - Desambiguação
  - Equivalência entre linguagens (incompatibilidades)

Google Translate

Exemplo: Mineração de Texto

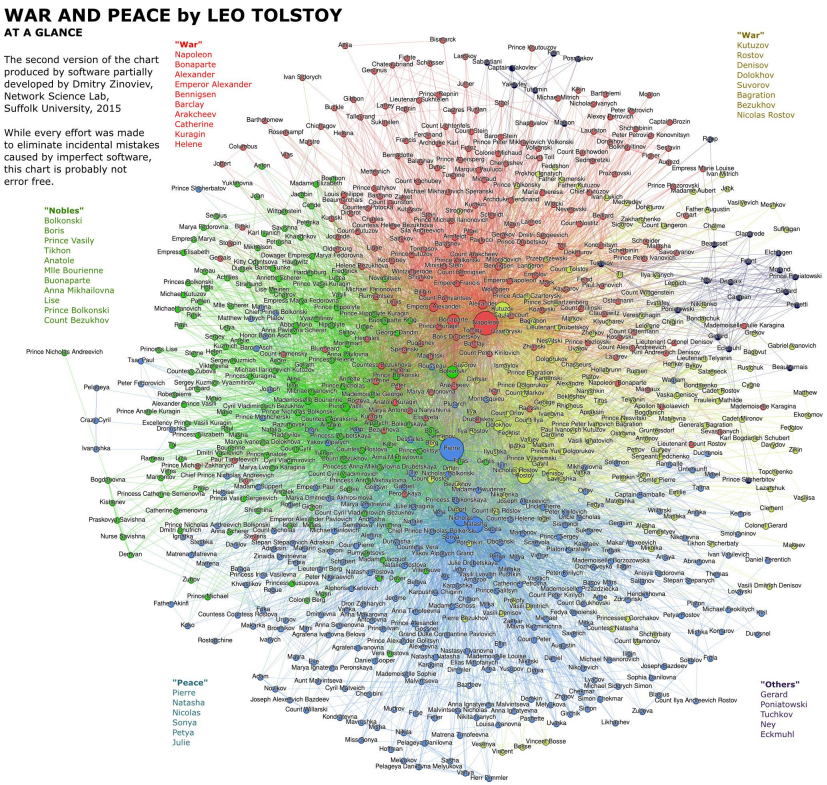
- Extração de novo conhecimento através da análise de textos
- Desafios linguísticos:
  - Sintaxe
  - Semântica
  - Extração de conceitos
  - *Sentiment analysis*
  - ...



Narrative Network of US Election 2012,
S Sudhahar, GA Veltri, N Cristianini

Exemplo: Mineração de Texto

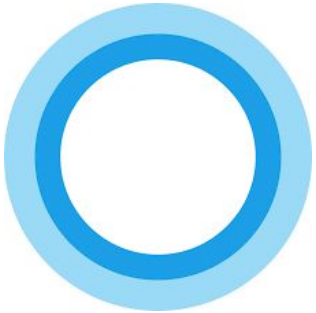
- Extração de novo conhecimento através da análise de textos
- Desafios linguísticos:
  - Sintaxe
  - Semântica
  - Extração de conceitos
  - *Sentiment analysis*
  - ...



Personagens da "Guerra e Paz", Dmitry Zinoviev

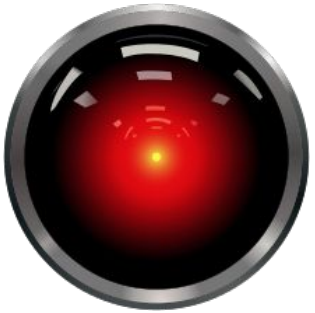
Exemplo: Sistemas de Diálogo

- Sistemas que comunicam em linguagem natural
- Problemas ao nível da linguagem
  - Reconhecimento de fala/voz
  - Compreensão de linguagem natural
  - Síntese de fala/voz



Cortana, Microsoft

Siri, Apple



HAL 9000, 2001: A Space Odyssey

E esta cadeira?

And this chair?

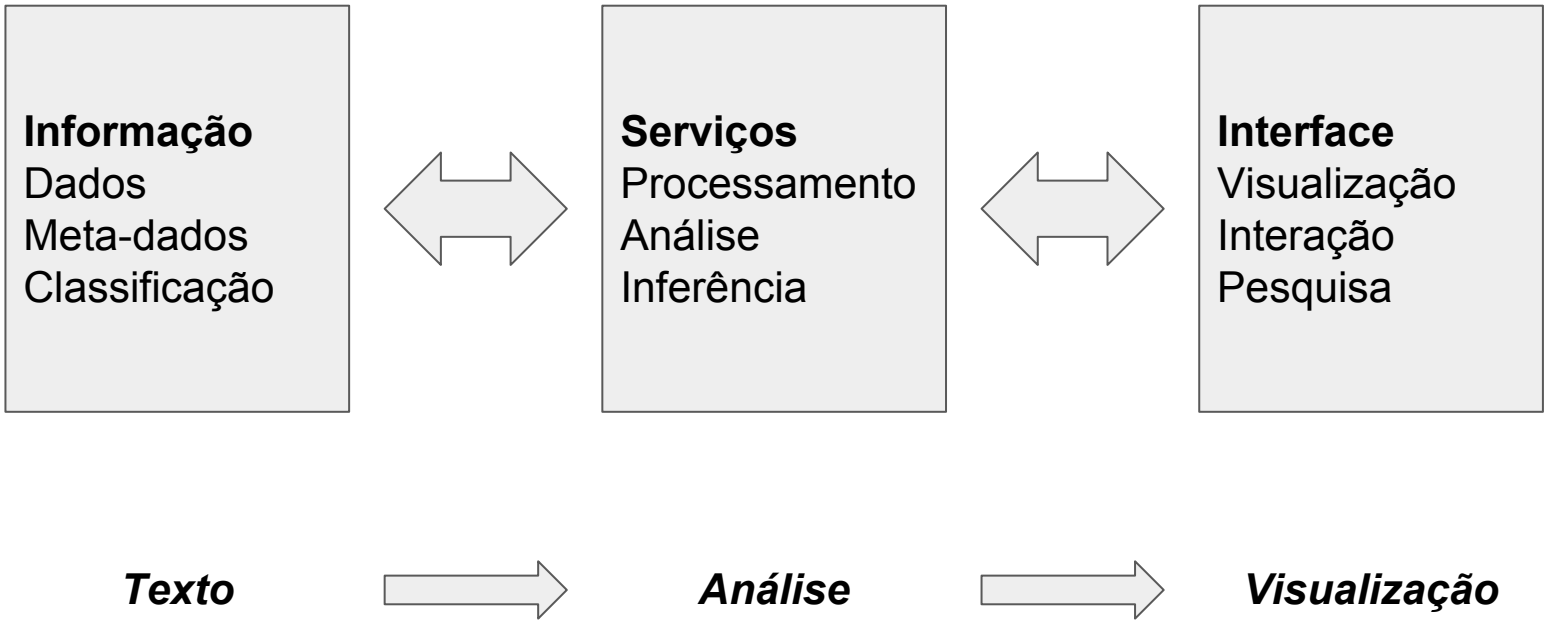
Google Translate

Aplicações de *Software*

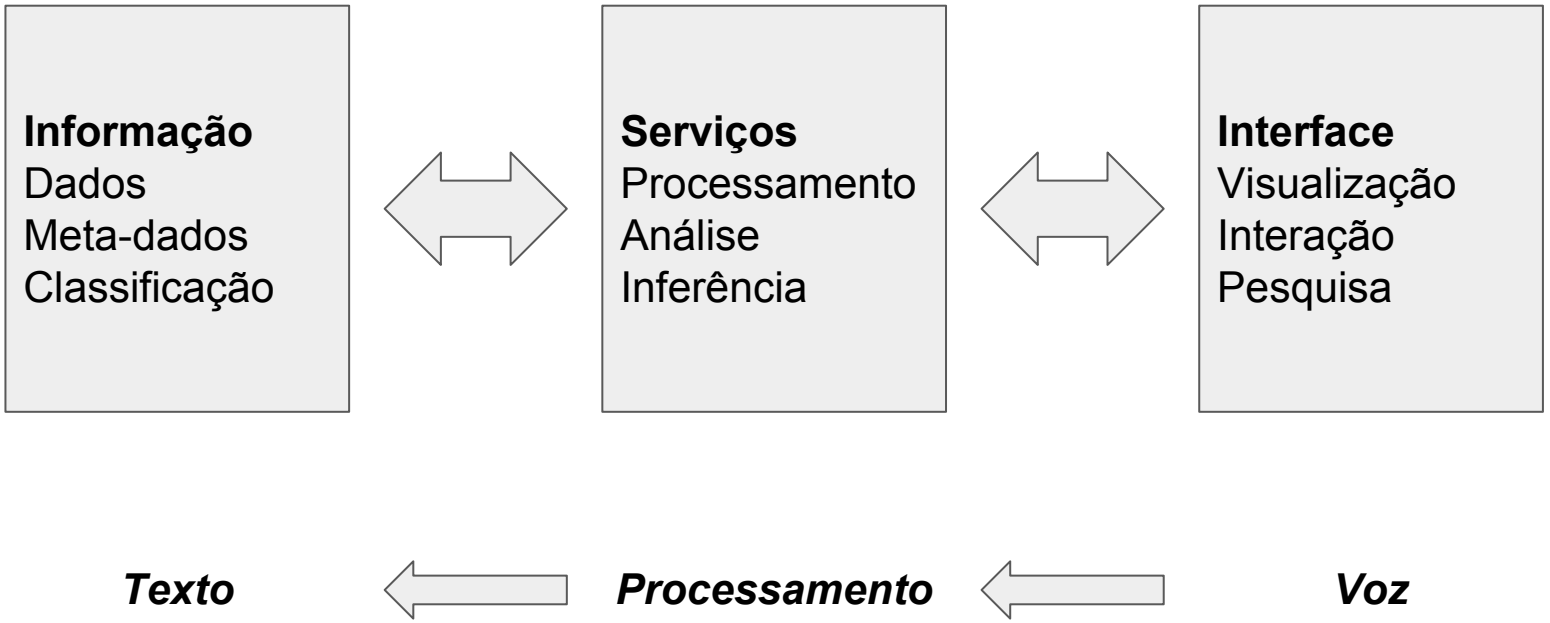
- Uma aplicação de *software* envolve sempre várias **camadas** com diferentes funcionalidades que passam informação entre si
  - *Dados:* representação da informação
  - *Serviços:* manipulação da informação
  - *Interface:* apresentação da informação
- Fases inter-dependentes em ambas as direções
- Permite construir funcionalidades **complexas** a partir de outros processos mais simples
- Geralmente aproveitam-se ferramentas desenvolvidas por **terceiros**

Aplicações de *Software*

Aplicações de *Software*



Aplicações de *Software*



Aplicações de *Software*

Texto

Desafios

”Within a computer natural language is unnatural.”

Alan Perlis

- A linguagem natural é intrinsecamente **ambígua**…
- E os computadores são intrinsecamente **exatos**!
- A informação tem que passar por um processo de digitalização para que seja computável

WATCHING THE UNICODE PEOPLE TRY TO GOVERN THE INFINITE CHAOS OF HUMAN LANGUAGE WITH CONSISTENT TECHNICAL STANDARDS IS LIKE WATCHING HIGHWAY ENGINEERS TRY TO STEER A RIVER USING TRAFFIC SIGNS.

XKCD

Desafios

“Que raiva ter esquecido o paiosinho! Enfim, acabou-se. Ao menos assentamos a teoria definitiva da existência.”

Os Maias, Eça de Queirós

- Impossível para uma máquina processar automaticamente
- Mas também impossível tornar toda a informação explícita!
- Fonética, morfológica, sintática, semântica, pragmática…
- Mesmo para o ser humano torna-se difícil lidar com toda esta informação

Desafios

“Não ouviste, que estás a fazer com esse pau, tomou o pai a perguntar, e o filho, sem levantar a vista da operação, respondeu, Estou a fazer uma tigela para quando o pai for velho e lhe tremerem as mãos...”

As Intermitências da Morte, José Saramago

- Impossível para uma máquina processar automaticamente
- Mas também impossível tornar toda a informação explícita!
- Fonética, morfológica, sintática, semântica, pragmática…
- Mesmo para o ser humano torna-se difícil lidar com toda esta informação

Linguagens de Anotação de Documentos

- É fundamental introduzir **estrutura**, criar modelos formais passíveis de serem computáveis
- No entanto, não queremos que os documentos se tornem **ininteligíveis** para humanos
- **Linguagens de anotação de documentos**!
- Introduzem informação adicional (**meta-dados**) nos documentos (**dados**)
- Anotações podem ser de natureza variada
- Vantagem acrescida: obriga-nos a raciocinar sobre a estrutura, formalizar questões que podiam passar despercebidas
- São o foco desta cadeira

Anotação de Documentos

Voyant Tools, Jane Austen Corpus

Objetivos

- Conceitos de informática básicos
  - Sistema de ficheiros
  - Processadores vs. editores de texto
  - *Encoding*
  - ...

Objetivos

- Linguagens de anotação de documentos
  - Separação aspecto/conteúdo
  - Formatos com estrutura explícita
  - LaTeX
  - HTML
  - XML e DTDs
  - Consultas básicas
  - Ontologias

Objetivos

- Ferramentas de trabalho colaborativo
  - Sistemas na “nuvem”
  - Controlo de versões
  - *Google Drive/Docs*
  - *ShareLaTeX*
  - ...

Disclaimer

- Não sou linguista (computacional)!
- Mas os conceitos tecnológicos fundamentais são suficientemente genéricos para qualquer domínio de aplicação
- Quanto a aplicações e ferramentas concretas, o plano é flexível

Take-home Message

- As tecnologias da informação têm um papel central em muitas áreas das humanidades, incluindo a linguística
- As técnicas vão desde a estruturação da informação à extração de novo conhecimento dessa informação
- A estruturação da informação é por si só desafiante devido à incompatibilidade entre a linguagem natural e modelos de computação
- Nesta cadeira vamos estudar tecnologias para facilitar essa estruturação através de linguagens de anotação de documentos
- No processo vamos explorar outros conceitos informáticos básicos

Questões Práticas

- Avaliação
  - 50% exame teórico
  - 50% exercícios práticos
- Guiões laboratoriais semanais, com 3 *milestones* para avaliações
- Grupos de trabalho de 3 elementos
- Aulas práticas: idealmente com computadores pessoais
- Contacto: Nuno Macedo, nfmmacedo@di.uminho.pt
- Atendimento: DI 2.07, ?????

Guião 0

- Criar conta *Gmail* para a comunicação na disciplina
- Aceder à pasta partilhada no *Google Drive*:

  https://drive.google.com/open?id=0B4c_AvViGhijanJib25ieHQtNnc
- Aceder ao documento partilhado "*Grupos*", preencher com dados dos elementos do grupo
- Preencher o formulário partilhado "*Ferramentas*" com ferramentas de utilização úteis no contexto do curso