II Seminário da Classificação Nacional de Atividades Econômicas - CNAE

Classificação Automática em CNAE

Desafios e Perspectivas

Diálogo Econômico do Brasil com o Mundo

Pesquisadores

Dr Alberto Fereira De Souza (DI/UFES)

Dr Elias Oliveira (DCI/UFES)

Dr Hannu Tapio Ahonen (DI/UFES)

Dr Felipe Maia Galvão França (COPPE/UFRJ)

Dr Priscila Machado Vieira Lima

Dr Eliana Zandonade (CCE/UFES)

AGENDA

1. O Problema e Motivação;
2. Estágio Atual da Pesquisa;
3. Próximos Passos;
4. Alguns Desafios à Frente.

O Problema e Motivação

Nosso problema é: *ler* o **objeto social** e classificá-lo em uma ou mais das **sub-classes** da tabela CNAE.

O Problema e Motivação

Nosso problema é: *ler* o **objeto social** e classificá-lo em uma ou mais das **sub-classes** da tabela CNAE.

Entretanto, muitas vezes, nem tudo é descrito no **objeto social**.

O Problema e Motivação

Nosso problema é: *ler* o **objeto social** e classificá-lo em uma ou mais das **sub-classes** da tabela CNAE.

Entretanto, muitas vezes, nem tudo é descrito no **objeto social**.

Resolver esse problema . . .

- Contribui com a desburocratização de processos;
- Facilita o monitoramento *online* dos setores;
- Racionaliza e precisa a fiscalização.

Calculando Documentos...

Nós seres humanos "pensamos", as máquinas "fazem contas"...

Calculando Documentos...

Nós seres humanos "pensamos", as máquinas "fazem contas"...

Vamos supor que tenhamos uma base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e queiramos saber quão similar q (um outro documento) é de um ou mais documentos em D.

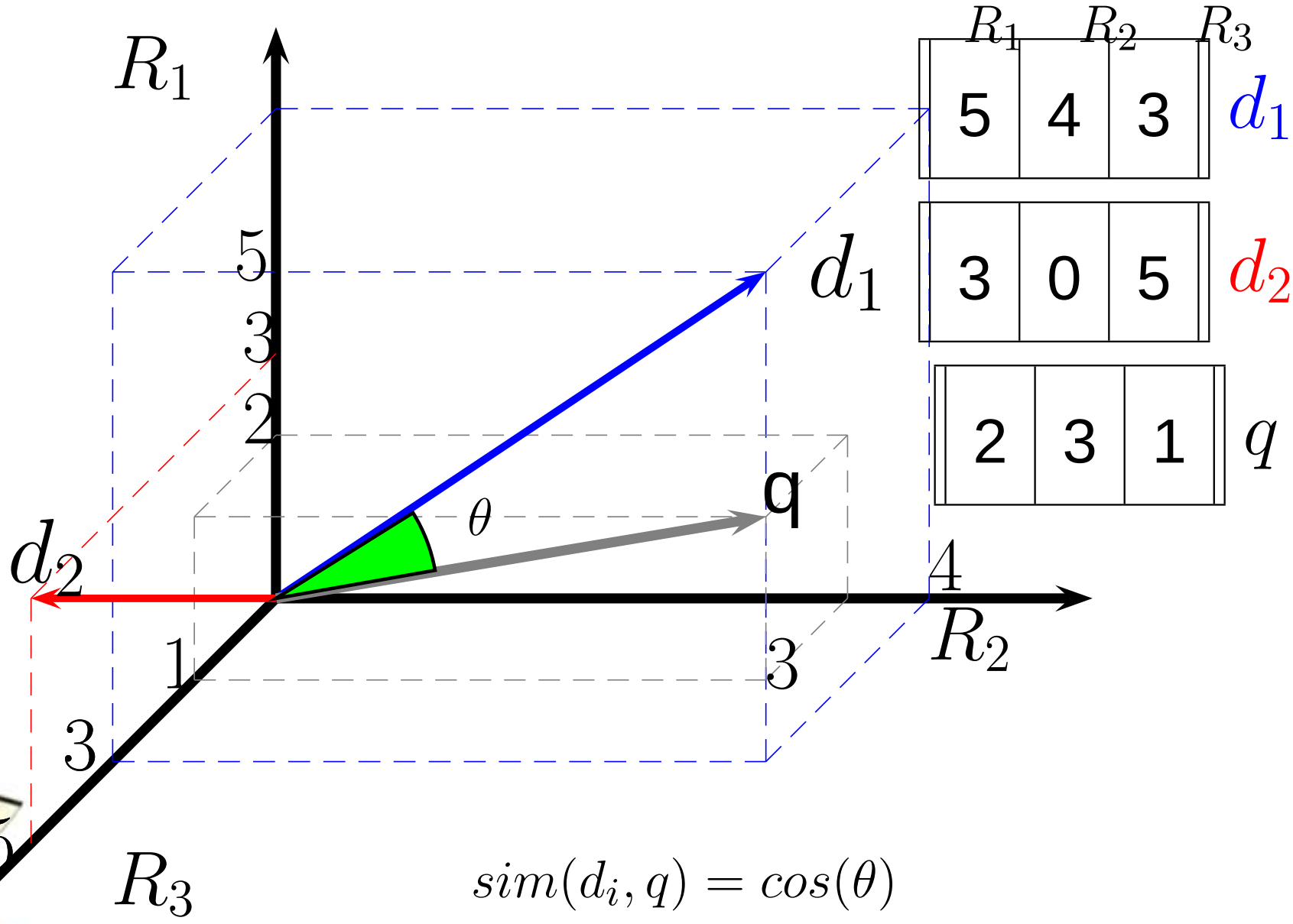
Calculando Documentos...

Nós seres humanos "pensamos", as máquinas "fazem contas"...

Vamos supor que tenhamos uma base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e queiramos saber quão similar q (um outro documento) é de um ou mais documentos em D.

Precisamos transformar o processo de classificação em um processo de contagem/cálculo...

Visualizando Documentos...



Estágio Atual da Pesquisa (I)

São três as técnicas prevista nessa pesquisa:

1. Modelo Vetorial;
2. Redes Neurais;
3. Redes Bayesianas.

Estágio Atual da Pesquisa (I)

Metodologia dos Experimentos

1. Extração das palavras da base de dados;
2. Filtragem de: *stopwords*, números, símbolos e *etc.*;
3. **Indexação** de cada documento em forma de um vetor...;
4. Cálculo da similaridade entre os documentos...;
5. Classificação dos documentos baseando-se na classe do mais similar;
6. Contabilização dos acertos de **pelo menos um código** do documento.

Estágio Atual da Pesquisa (II)

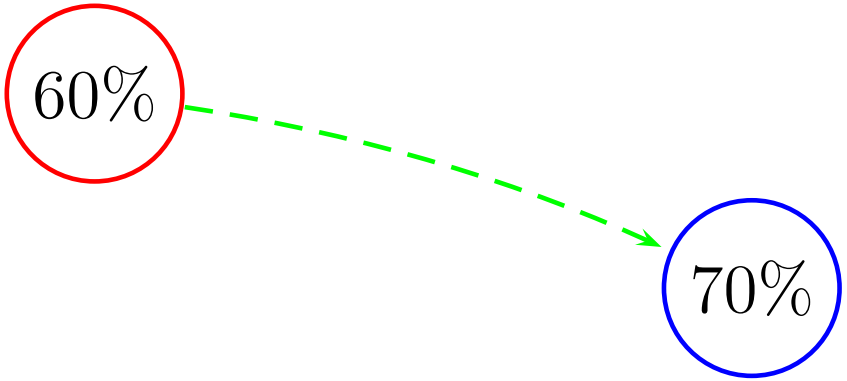
Resultados obtidos até o momento:

Modelo Vetorial	Redes Neurais
63.36%	67.03%

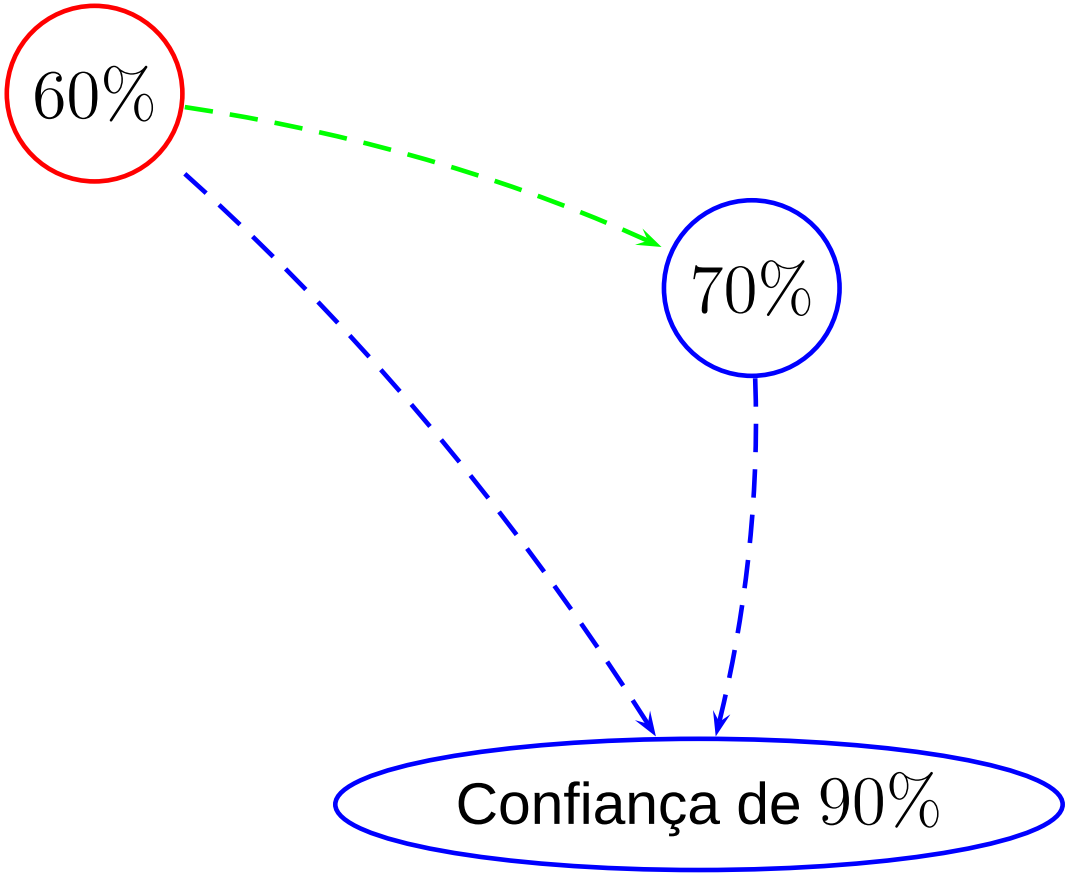
Próximos Passos

60%

Próximos Passos



Próximos Passos



Alguns Desafios à Frente

- Avaliação dos resultados com outras métricas;

Alguns Desafios à Frente

- Avaliação dos resultados com outras métricas;
- Experimentação com outras bases de dados;

Alguns Desafios à Frente

- Avaliação dos resultados com outras métricas;
- Experimentação com outras bases de dados;
- Utilização de uma base de dados **corretamente** classificada;

Alguns Desafios à Frente

- Avaliação dos resultados com outras métricas;
- Experimentação com outras bases de dados;
- Utilização de uma base de dados **corretamente** classificada;
- Estudar especificidades de **áreas críticas**;

Alguns Desafios à Frente

- Avaliação dos resultados com outras métricas;
- Experimentação com outras bases de dados;
- Utilização de uma base de dados **corretamente** classificada;
- Estudar especificidades de **áreas críticas**;
- Estudar soluções de *software* e *hardware* **escaláveis**.

OBRIGADO!

Referências

De Souza, A. F. et al. Automated Free Text Classification of Economic Activities using VG-RAM Weightless Neural Networks. In: *7th International Conference on Intelligent Systems Design and Applications*. Rio de Janeiro: [s.n.], 2007.

OLIVEIRA, E. et al. Intelligent Classification of Economic Activities from Free Text Descriptions. In: *5^{o} Workshop em Tecnologia da Informação e da Linguagem Humana*. Rio de Janeiro: [s.n.], 2007.

elias@inf.ufes.br