

Classificação automática de teses e dissertações da área da Ciência da Informação sob a ótica dos Grupos de Trabalho da Ancib

André Fabiano Dyck
andre.dyck@ufsc.br

Moisés Lima Dutra
Moises.dutra@ufsc.br

Angel Freddy Godoy Viera
a.godoy@ufsc.br



Apresentação

Graduação e Mestrado em Ciência da Computação

Analista de TI (SeTIC/UFSC)

Aluno de doutorado - turma 2019 (PGCIN/UFSC)



Motivação

Por que classificar teses e dissertações?



Motivação

Localização manual pode ser trabalhosa e demorada



Motivação

Considerando que os eixos temáticos de pesquisa dos PPG em CI possuem alinhamento com os GTs Ancib



Fonte: <http://gtancib.fci.unb.br/>

Motivação

Considerando que os eixos temáticos de pesquisa dos PPG em CI possuem alinhamento com os GTs Ancib



Organização e preservação do conhecimento

Profissionais da informação, competência em informação e publicação científica

Informação e tecnologia

Gestão da informação e do conhecimento

GT 02

Organização e Representação do Conhecimento

GT 04

Gestão da Informação e do Conhecimento

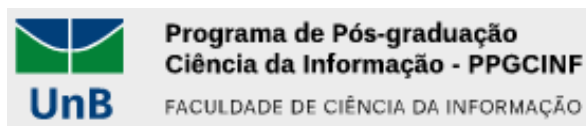
GT 08

Informação e Tecnologia

Fontes: <https://pgcin.paginas.ufsc.br/linhas-de-pesquisa/>
<http://gtancib.fci.unb.br>

Motivação

Considerando que os eixos temáticos de pesquisa dos PPG em CI possuem alinhamento com os GTs Ancib



Organização da informação

Comunicação e mediação da informação

GT 02

Organização e Representação
do Conhecimento

GT 03

Mediação, Circulação e
Apropriação da Informação

GT 07

Produção e Comunicação da
Informação em Ciência,
Tecnologia & Inovação

Fontes: <http://ppgcinf.fci.unb.br/pt/programa/sobre>
<http://gtancib.fci.unb.br>

Motivação

Considerando que os eixos temáticos de pesquisa dos PPG em CI possuem alinhamento com os GTs Ancib



Informação e Tecnologia

Produção e Organização da Informação

Gestão, Mediação e Uso da Informação

GT 02

Organização e Representação
do Conhecimento

GT 03

Mediação, Circulação e
Apropriação da Informação

GT 07

Produção e Comunicação da
Informação em Ciência,
Tecnologia & Inovação

GT 04

Gestão da Informação e do
Conhecimento

GT 08

Informação e Tecnologia

Fontes: <https://www.marilia.unesp.br/#!/pos-graduacao/mestrado-e-doutorado/ciencia-da-informacao/linhas-de-pesquisa/>
<http://gtancib.fci.unb.br>

Motivação

Fazer um trabalho direcionado, que serve como um experimento

Depois, adequar minhar tese



Perguntas de pesquisa

Qual o alinhamento das teses e dissertações da área de CI com os GTs da Ancib?

Quais são as temáticas mais trabalhadas pelos PPG em CI?

Como localizar as teses e dissertações alinhadas a um GT específico?

...



Procedimentos metodológicos

Utilizamos o modelo Saco-de-Palavras (*Bag-of-Words – BoW*)
Utilizando apenas as palavras e suas combinações (n-gramas)
Sem considerar a semântica

unigrama (n-grama=1): "INFORMAÇÃO"

bigrama (n-grama=2): "BIBLIOTECAS UNIVERSITÁRIAS"

trigrama (n-grama=3): "CIÊNCIA DA INFORMAÇÃO"



Procedimentos metodológicos

Piloto: PGCIN/UFSC (<https://repositorio.ufsc.br/>)

Coleta e tabulação entre 26/08/2019 e 09/09/2019

223 documentos (teses e dissertações)

185 resumos

11 ementas GTs Ancib (<http://gtancib.fci.unb.br/>)

Procedimentos metodológicos



The screenshot shows the homepage of the UFSC Institutional Repository (Repositório Institucional da UFSC). The browser address bar displays "repositorio.ufsc.br". The page features a search bar, navigation links for "All of DSpace", "My Account", and "Discover". The "Discover" section lists author communities, with a blue arrow pointing to the "Trabalhos Acadêmicos" entry.

repositorio.ufsc.br

Login

RI UFSC REPOSITÓRIO INSTITUCIONAL UFSC

DSpace Home

Search DSpace

Browse

All of DSpace

- [Communities & Collections](#)
- [By Issue Date](#)
- [Authors](#)
- [Titles](#)
- [Subjects](#)

My Account

[Login](#)

Discover

Author

- [Tempo Editorial \(19310\)](#)
- [Anacleto, Waldemar \(1419\)](#)
- [Agência de](#)

Repositório Institucional da UFSC

Bem-vindo ao Repositório Institucional da UFSC.

O Repositório Institucional (RI) da Universidade Federal de Santa Catarina (UFSC) tem como missão: armazenar, preservar, divulgar e oferecer acesso à produção científica e institucional da UFSC.

Possui como objetivos: contribuir para o aumento da visibilidade da produção científica da UFSC; preservar a memória intelectual da Universidade; reunir em um único local virtual e de forma permanente a produção científica e institucional; disponibilizar o livre acesso aos conteúdos digitais; ampliar e facilitar o acesso à produção científica de uma forma geral.

O Repositório Institucional da UFSC é uma iniciativa de acesso aberto e gratuito. Possui como licença padrão a CC BY-NC 3.0 BR. Mais informações: <http://www.repositorio.ufsc.br/licenca/>.

Regras, manuais e outras informações sobre o RI podem ser encontradas em: <http://www.repositorio.ufsc.br>.

Communities in DSpace

Select a community to browse its collections.

- [Acervos \[57224\]](#)
- [Teses e Dissertações \[37181\]](#)
- [Trabalhos Acadêmicos \[24694\]](#)

Procedimentos metodológicos

The screenshot shows the UFSC Institutional Repository website. The browser address bar displays 'repositorio.ufsc.br/handle/123456789/74645'. The page header includes the UFSC logo and 'REPOSITÓRIO INSTITUCIONAL UFSC'. Below the header, there is a navigation bar with 'DSpace Home' and 'Teses e Dissertações'. The main content area is titled 'Teses e Dissertações' and contains a search bar, a description of the community, and a list of collections. A blue arrow points to the 'Programa de Pós-Graduação em Ciência da Computação' collection.

repositorio.ufsc.br/handle/123456789/74645

Login

RI UFSC REPOSITÓRIO INSTITUCIONAL UFSC

DSpace Home » Teses e Dissertações

Search DSpace

Go

Search DSpace
 This Community

Browse

All of DSpace
[Communities & Collections](#)
[By Issue Date](#)
[Authors](#)
[Titles](#)
[Subjects](#)

This Community
[By Issue Date](#)
[Authors](#)
[Titles](#)
[Subjects](#)

My Account
[Login](#)

Discover

Author
[Silva, Edson Luiz da \(6\)](#)
[Ernandorena, Paulo Renato \(5\)](#)


Teses e Dissertações

Search within this community and its collections: Go

Nesta comunidade são apresentadas as coleções de teses e dissertações dos Programas de Pós-Graduação da Universidade Federal de Santa Catarina.

Collections in this community

- [Programa de Pós-Graduação em Administração \[706\]](#)
- [Programa de Pós-Graduação em Administração Universitária \(Mestrado Profissional\) \[172\]](#)
- [Programa de Pós-Graduação em Agroecossistemas \[354\]](#)
- [Programa de Pós-Graduação em Agroecossistemas \(Mestrado Profissional\) \[53\]](#)
- [Programa de Pós-Graduação em Antropologia Social \[380\]](#)
- [Programa de Pós-Graduação em Aquicultura \[425\]](#)
- [Programa de Pós-Graduação em Arquitetura e Urbanismo \[280\]](#)
- [Programa de Pós-Graduação em Assistência Farmacêutica \[7\]](#)
- [Programa de Pós-Graduação em Biologia Celular e do Desenvolvimento \[91\]](#)
- [Programa de Pós-Graduação em Biologia de Fungos, Algas e Plantas \[80\]](#)
- [Programa de Pós-Graduação em Biologia Vegetal \[137\]](#)
- [Programa de Pós-Graduação em Bioquímica \[128\]](#)
- [Programa de Pós-Graduação em Biotecnologia \[174\]](#)
- [Programa de Pós-Graduação em Biotecnologia e Biociências \[108\]](#)
- [Programa de Pós-Graduação em Ciência da Computação \[878\]](#)
- [Programa de Pós-Graduação em Ciência da Informação \[223\]](#)
- [Programa de Pós-Graduação em Ciência dos Alimentos \[303\]](#)



Procedimentos metodológicos

The image shows a web browser window displaying the website 'gtancib.fci.unb.br'. The page title is 'Fórum de Coordenadores de Grupo de Trabalho da Ancib'. The website features a navigation menu with 'Início', 'Apresentação', 'Documentos', and 'Grupos de Trabalho Desativados'. A search bar is located in the top right corner. The main content area is titled 'Sejam Bem-Vindos!' and lists 11 research groups (GT 01 to GT 11) in blue boxes, each with a specific topic.

GT	Descrição
GT 01	Estudos Históricos e Epistemológicos da Ciência da Informação
GT 02	Organização e Representação do Conhecimento
GT 03	Mediação, Circulação e Apropriação da Informação
GT 04	Gestão da Informação e do Conhecimento
GT 05	Política e Economia da Informação
GT 06	Informação, Educação e Trabalho
GT 07	Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação
GT 08	Informação e Tecnologia
GT 09	Museu, Patrimônio e Informação
GT 10	Informação e Memória
GT 11	Informação & Saúde

Procedimentos metodológicos

Criação dos dicionários de termos

Limpeza dos dados para unigramas, bigramas, trigramas:

Remoção de pontuação

Transformação das palavras em MAIÚSCULAS

Limpeza de dados, **apenas**, para unigramas:

Remoção de palavras sem valor semânticos (*stop words*)

(de, do, da, dos, das, a, as, o, os, que, em, para, na, nas, ...)

Procedimentos metodológicos

Com os dicionários de termos prontos (criados e limpos)

Tanto para os Resumos como para as Ementas

Contabilizamos a frequência das ocorrências de cada termo

Partimos para a classificação automática



Arquitetura proposta

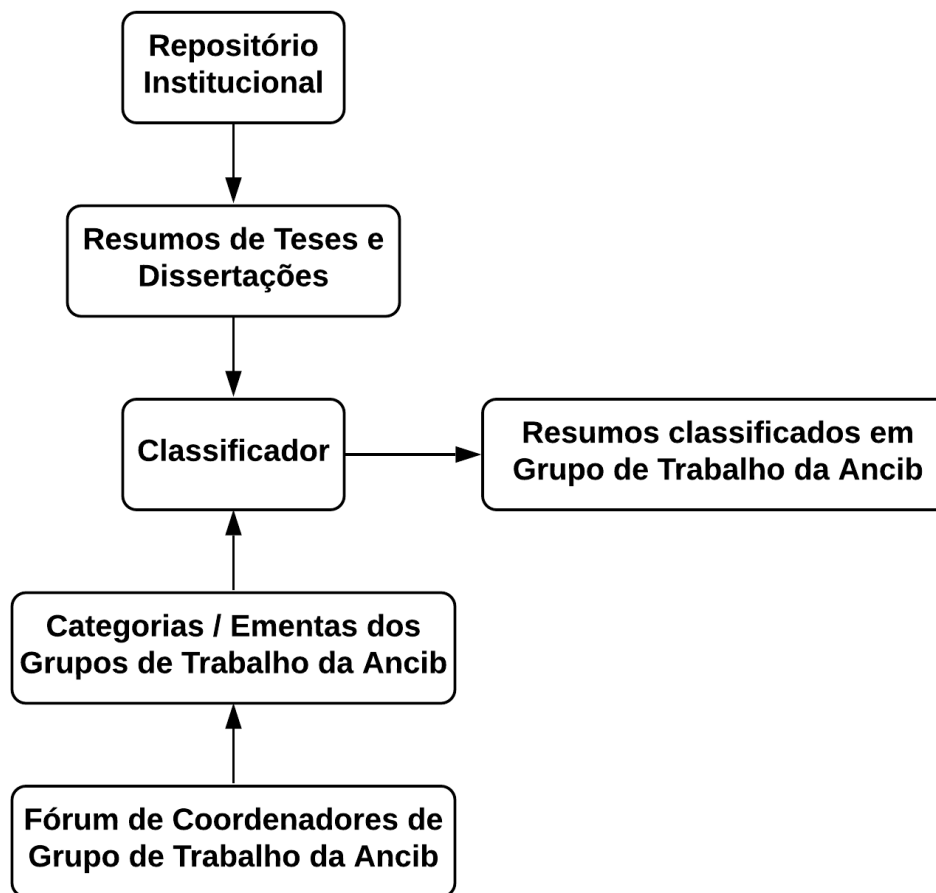
Usamos a técnica de aprendizagem de máquina supervisionada

Classificação textual (*text classification*)

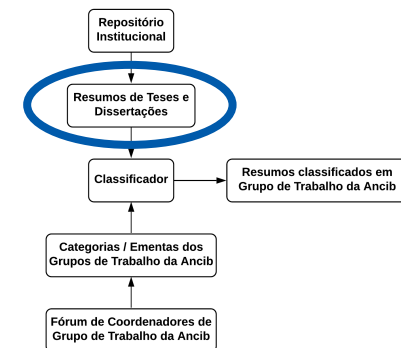
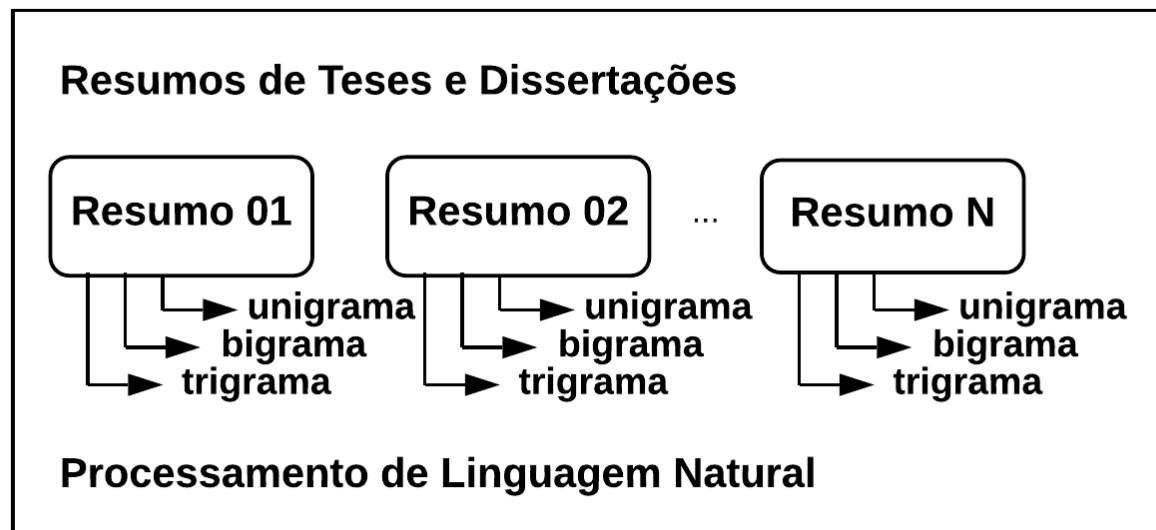
As categorias (Ementas dos GTs da Ancib) foram humana e previamente definidas



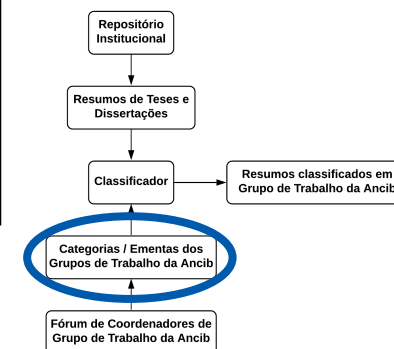
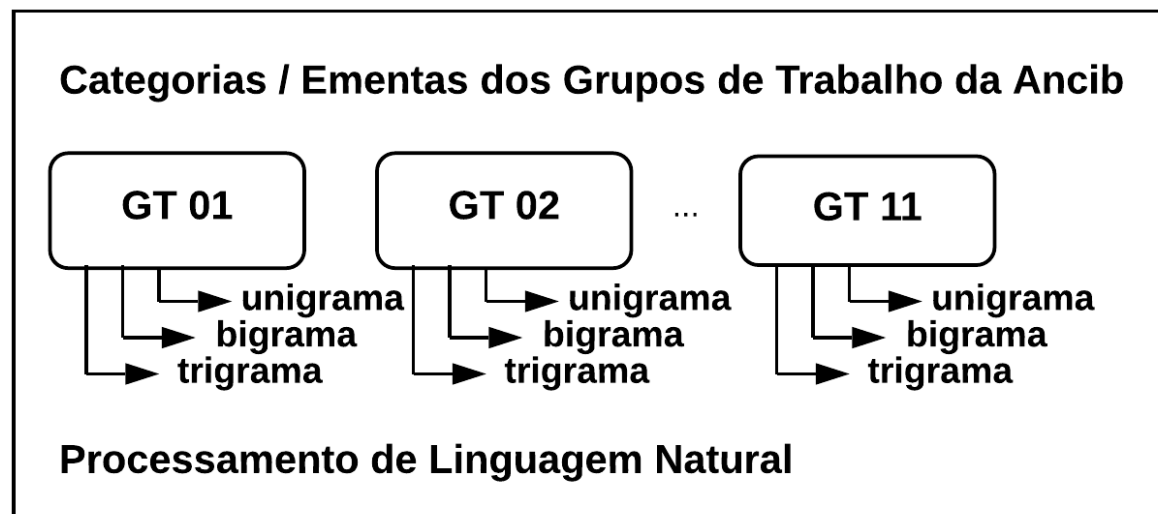
Arquitetura



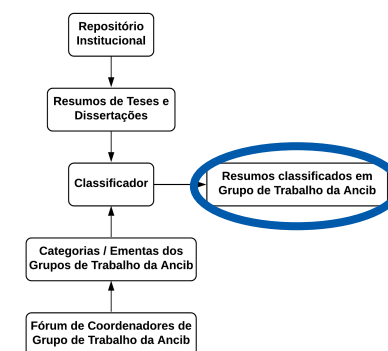
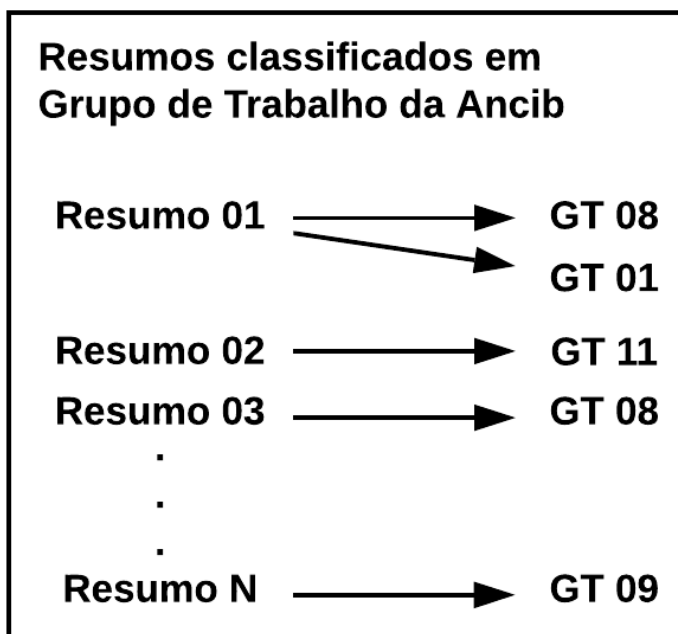
Arquitetura



Arquitetura

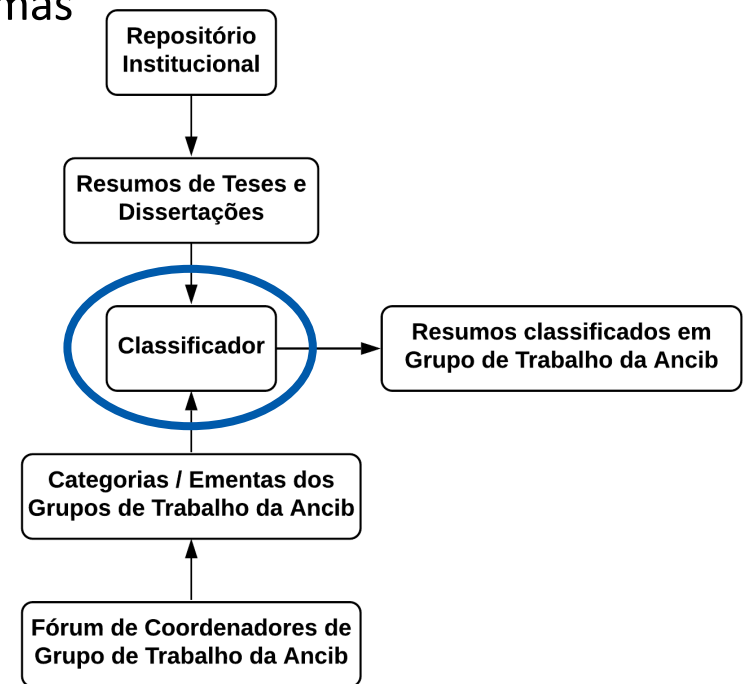


Arquitetura



Classificador

1. Comparação entre unigramas, bigramas e trigramas
Comparação de cada Resumo com cada Ementa



Classificador

1. Comparação entre unigramas, bigramas e trigramas

Comparação de cada Resumo com cada Ementa

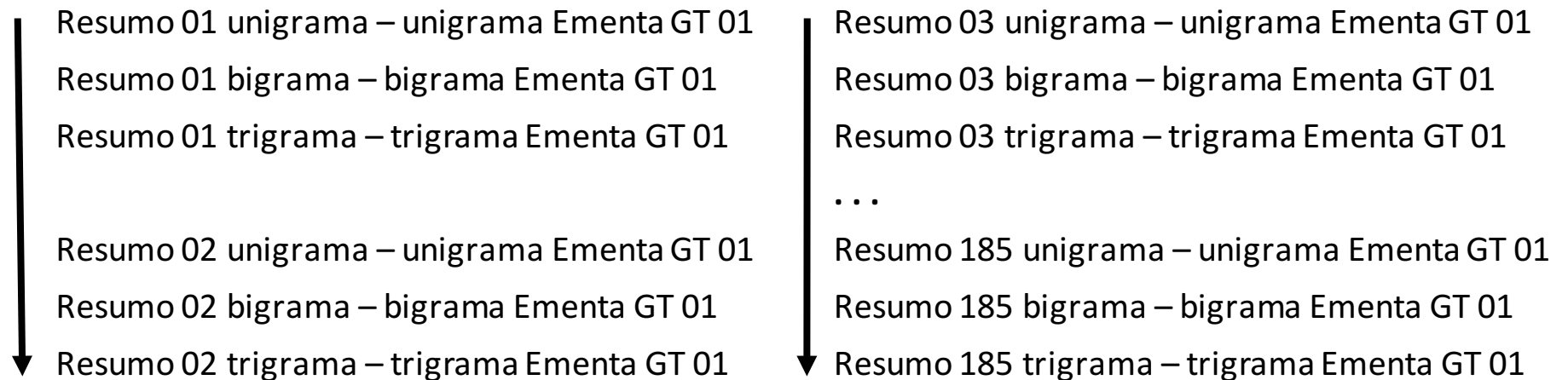
↓
Resumo 01 – Ementa GT 01
Resumo 02 – Ementa GT 01
Resumo 03 – Ementa GT 01
...
Resumo 185 – Ementa GT 01

↓
Resumo 01 – Ementa GT 02
Resumo 02 – Ementa GT 02
Resumo 03 – Ementa GT 02
...
Resumo 185 – Ementa GT 02
...
Resumo 185 – Ementa GT 11

Classificador

1. Comparação entre unigramas, bigramas e trigramas

Comparação de cada Resumo com cada Ementa



Classificador

1. Comparação entre unigramas, bigramas e trigramas

Comparação de cada Resumo com cada Ementa

Resumo 01 unigrama – unigrama Ementa GT **02**

Resumo 01 bigrama – bigrama Ementa GT **02**

Resumo 01 trigrama – trigrama Ementa GT **02**

Resumo 02 unigrama – unigrama Ementa GT **02**

Resumo 02 bigrama – bigrama Ementa GT **02**

Resumo 02 trigrama – trigrama Ementa GT **02**

Resumo 03 unigrama – unigrama Ementa GT **02**

Resumo 03 bigrama – bigrama Ementa GT **02**

Resumo 03 trigrama – trigrama Ementa GT **02**

...

Resumo 185 unigrama – unigrama Ementa GT **02**

Resumo 185 bigrama – bigrama Ementa GT **02**

Resumo 185 trigrama – trigrama Ementa GT **02**

Classificador

1. Comparação entre unigramas, bigramas e trigramas

Comparação de cada Resumo com cada Ementa

Resumo 01 unigrama – unigrama Ementa GT **11**

Resumo 01 bigrama – bigrama Ementa GT **11**

Resumo 01 trigrama – trigrama Ementa GT **11**

Resumo 02 unigrama – unigrama Ementa GT **11**

Resumo 02 bigrama – bigrama Ementa GT **11**

Resumo 02 trigrama – trigrama Ementa GT **11**

Resumo 03 unigrama – unigrama Ementa GT **11**

Resumo 03 bigrama – bigrama Ementa GT **11**

Resumo 03 trigrama – trigrama Ementa GT **11**

...

Resumo 185 unigrama – unigrama Ementa GT **11**

Resumo 185 bigrama – bigrama Ementa GT **11**

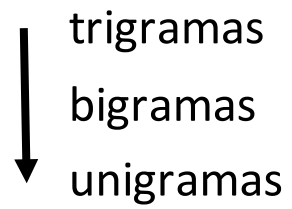
Resumo 185 trigrama – trigrama Ementa GT **11**

Classificador

2. Geração de índices resultantes da comparação, entre os unigramas, bigramas e trigramas de um Resumo com uma Ementa

Classificador

3. Atribuição de pesos para estes índices

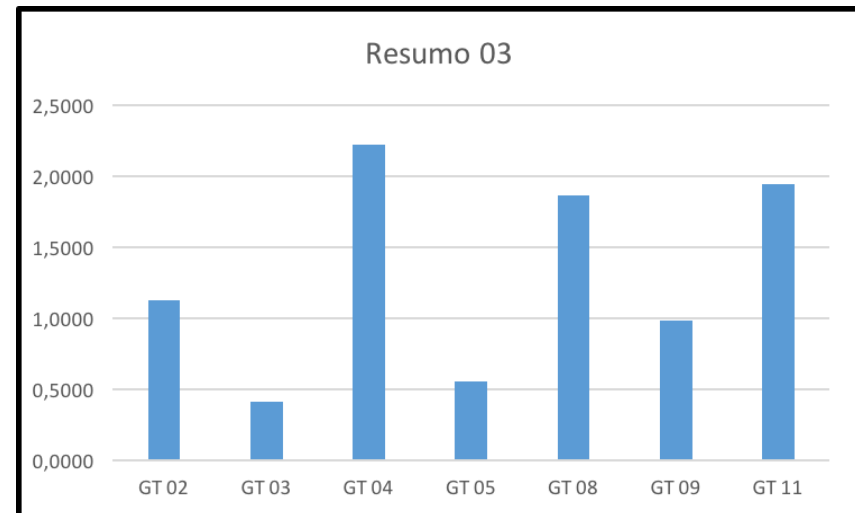
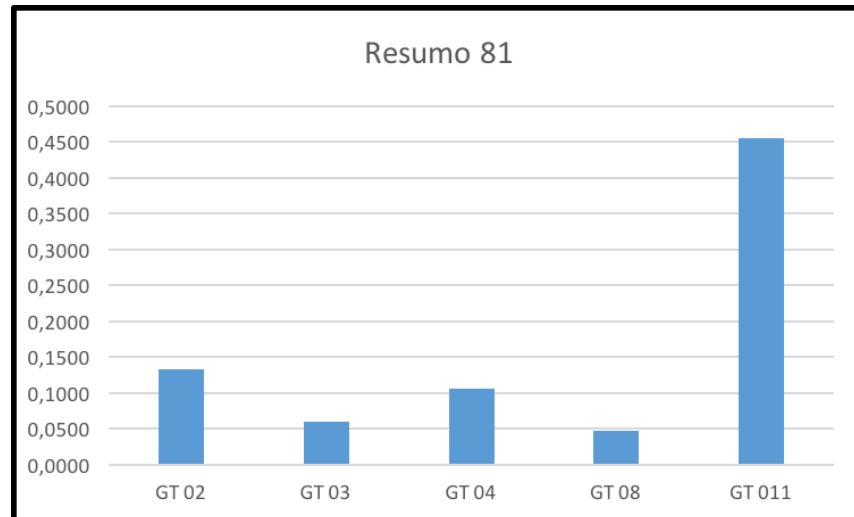


Classificador

4. Criação de uma fórmula com os pesos e os índices dos trigramas, bigramas e unigramas

Aplicação desta fórmula nos resultados da comparação para definir a probabilidade de pertencimento de um resumo a um determinado GT

Alguns resultados



Próximas etapas

Construção e validação do código

Aumentar o número de n-gramas

Testar o modelo em um método não-supervisionado (Clusterização)

Usando volumes maiores de dados

Substituir BoW por técnica que preserva semântica (*Word Embeddings*)

Utilizar teses e dissertações dos 23 PPG em CI

Classificação automática de teses e dissertações da área da Ciência da Informação sob a ótica dos Grupos de Trabalho da Ancib

André Fabiano Dyck
andre.dyck@ufsc.br

Moisés Lima Dutra
Moises.dutra@ufsc.br

Angel Freddy Godoy Viera
a.godoy@ufsc.br

