

UNIVERSIDADE FEDERAL DE SANTA CATARINA – UFSC
CENTRO DE CIÊNCIAS DA EDUCAÇÃO – CED
DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO – CIN

CIN 7502 - PLANO DE ENSINO (2019.1)

1 IDENTIFICAÇÃO

DISCIPLINA: CIN 7502 – Mineração de Texto
CARGA HORÁRIA: 36h semestrais / 2h semanais
PROFESSOR: Moisés Lima Dutra (moises.dutra@ufsc.br)
OFERTA: Obrigatória para a 5ª. Fase do Curso de Ciência da Informação
HORA E LOCAL: Terças-feiras, das 18h30 às 20h10, LABINFOR

2 EMENTA

Técnicas de mineração de texto. Mineração de conteúdo Web. Pré-processamento de texto: palavras proibidas, radicalização, substantivação, reconhecimento de entidades nomeadas. Processamento de texto: categorização, agrupamento, sumarização, extração da informação. Pós-processamento: métricas de avaliação.

3 OBJETIVOS

3.1 Objetivo Geral

Desenvolver a capacidade de extrair informações importantes de textos, por meio da aplicação de padrões e técnicas de mineração de dados textuais.

3.2 Objetivos Específicos

- 3.2.1. Conhecer as técnicas de mineração de texto.
- 3.2.2. Preparar os dados textuais para a mineração.
- 3.2.3. Aplicar técnicas de mineração de texto em conjuntos específicos de dados textuais.
- 3.2.4. Avaliar os resultados minerados e pós-processados.
- 3.2.5. Capacitar-se na execução da mineração de texto nos mais diversos cenários de aplicação.

4 CONTEÚDO PROGRAMÁTICO

4.1 Introdução à Mineração de Texto

- 4.1.1 Definição e Evolução
- 4.1.2 Tipos de Dados
- 4.1.3 Padrões de Representação
- 4.1.4 KDT (Knowledge Discovery from Text)

4.2 Desenvolvimento da Mineração

- 4.2.1 Preparação dos Dados
- 4.2.2 Análise Léxica
- 4.2.3 Bag of Words
- 4.2.4 Modelo de Espaço Vetorial
- 4.2.5 Cálculo de Relevância (TF-IDF)
- 4.2.6 Recall (revocação)
- 4.2.7 Precision (precisão)

4.3 Mineração de Dados Textuais na Web

4.3.1 Web Crawlers

4.3.2 Processamento de Linguagem Natural

4.3.3 Análise de Redes Sociais

4.3.4 Mineração de Opiniões e Análise de Sentimentos

4.4 Machine Learning para Textos

4.4.1 Aprendizagem Supervisionada para Textos

4.4.2 Aprendizagem Não-Supervisionada para Textos

4.4.3 Classificação de Textos

4.4.4 Ordenamento por Relevância

5 METODOLOGIA

Aulas expositivas e atividades práticas de laboratório. Desenvolvimento de um projeto final de semestre.

O Moodle (<http://moodle.ufsc.br/>) será utilizado como instrumento de suporte da disciplina, devendo o mesmo, portanto, ser **consultado todas as semanas** pelos alunos.

6 AVALIAÇÃO

Conforme a Resolução do Conselho Universitário 017/Cun/97, de 30/09/97:

- ❖ A frequência mínima obrigatória é de 75% das aulas.
- ❖ **Em caso de falta em dia de avaliação (prova ou apresentação de trabalho)**, encaminhar justificativa e pedido formal à Chefia do Departamento de Ciência da Informação, no prazo de três (3) dias úteis.

Avaliações:

- Serão realizadas através do desenvolvimento de um projeto final de semestre, cuja apreciação progressiva se dará em quatro etapas: Etapa 1 (**E1**), Etapa 2 (**E2**), Etapa 3(**E3**) e Etapa Final (**EF**).

A média final (**MF**) será calculada conforme a seguinte fórmula:

$$\mathbf{MF} = 0.2(\mathbf{E1}) + 0.2(\mathbf{E2}) + 0.2(\mathbf{E3}) + 0.4(\mathbf{EF})$$

- Arredondamentos de notas só serão feitos com a Média Final (**MF**).
- ❖ O aluno que obtiver a **MF entre 3,0 e 5,5** e frequência suficiente poderá, ao final do semestre, realizar uma prova de recuperação de todo o conteúdo.
- ❖ **EM CASO DE RECUPERAÇÃO**, a nota final do semestre (**NFS**) será calculada a partir da **média simples** entre a média final (**MF**) obtida durante o semestre e a nota obtida na prova de recuperação (**PR**).

$$\mathbf{NFS} = (\mathbf{MF} + \mathbf{PR}) / 2$$

7 BIBLIOGRAFIA

Básica:

BENGFORT, Benjamin; BILBRO, Rebecca; OJEDA, Tony. **Applied Text Analysis with Python: Enabling Language Aware Data Products with Machine Learning**. Sebastopol, Ca: O'reilly, 2017.

SILGE, Julia; ROBINSON, David. **Text Mining with R: Tidy Approach**. Sebastopol, Ca: O'reilly, 2017. 194 p.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What you need to know about data mining and data-analytic Thinking**. Sebastopol (USA): O'Reilly Media, Inc., 2013. ISBN 978-1-449-36132-7.

WEISS, Sholom M.; INDURKHIA, Nitin; ZHANG, Tong. **Fundamentals of Predictive Text Mining**. New York: Springer, 2010. 226 p. (Texts in Computer Science).

Complementar:

BRZUSTOWICZ, Michael R. **Data Science with Java: Practical Methods for Scientists and Engineers**. Sebastopol, Ca: O'reilly, 2017. 236 p.

CAFARELLA, M. J.; HALEVY, A.; MADHAVAN, J. Structured Data on the Web. **Communications of the ACM**, vol. 54, no. 2, Fevereiro 2011.

FELDMAN, Ronen; SANGER, James. **The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data**. Cambridge, Ma: University Press, 2007. 410 p. Disponível em: <https://wtlab.um.ac.ir/images/e-library/text_mining/The_Text_Mining_HandBook.pdf>. Acesso em: 22 set. 2017.

FINLAY, S. **Predictive Analytics, Data Mining, and Big Data: Myths, Misconceptions and Methods**. Hampshire (UK): Palgrave Macmillan, 2014. 248 p. ISBN 978-1-137-37927-6.

FOREMAN, John W. **Data Smart: Usando Data Science para Transformar Informação em Insight**. Rio de Janeiro: Alta Books, 2016. 425 p.

GRIGORIEV, Alex. **Mastering Java for Data Science: Building data science applications in Java**. Birmingham (uk): Packt, 2017. 348 p.

HETLAND, Magnus Lie. **Beginning Python: From Novice to Professional**. Berkeley, CA: Apress, Inc., 2005. ISBN 9781430200727 Disponível em: <<http://dx.doi.org/10.1007/978-1-4302-0072-7>>. Acesso em: 08 jun.2017.

KAMATH, Dr. Uday; CHOPPELLA, Krishna. **Mastering Java Machine Learning: Mastering and implementing advanced techniques in machine learning**. Birmingham (uk): Packt, 2017. 556 p.

KAO, Anne; POTEET, Stephen R. (Ed.). **Natural Language Processing and Text Mining**. London: Springer-verlag, 2007. 265 p.

LANGTANGEN, Hans Petter. **Python Scripting for Computational Science**. Third Edition. Berlin: Springer-Verlag Berlin Heidelberg, 2008. (Texts in Computational Science and Engineering, 1611-0994; 3).

LUCENE. **Apache Lucene**. 2017. Disponível em <<https://lucene.apache.org/>>. Acesso em: 22 set. 2017.

SHROFF, G. **The Intelligent Web: Search, smart algorithms, and big data**. Oxford (UK): Oxford University Press, 2014. ISBN: 978-0199646715.

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à Mineração de Dados: Com Aplicações em R**. Rio de Janeiro: Elsevier, 2016. 277 p. (SBC).

W3C, 2013a. **Data Activity**. W3C, 2017. Disponível em: <<http://www.w3.org/2013/data/>>. Acesso em: 22 set. 2017.

WORDNETPT. **WordNet.PT**. 2017. Disponível em <<http://wordnet.pt/>>. Acesso em: 22 set. 2017.

8 CRONOGRAMA *		
AULA	DATA	CONTEÚDO
1	12/03	Apresentação da Disciplina. Introdução à Mineração de Texto.
2	19/03	Expressões Regulares.
3	26/03	Expressões Regulares.
4	02/04	Limpeza e Higiênização Textual.
5	09/04	(E1): Etapa 1 do Projeto Final de Semestre. Processamento de Linguagem Natural.
6	16/04	Web Scraping. Processamento de Linguagem Natural.
7	23/04	Web Scraping. Processamento de Linguagem Natural.
8	30/04	Web Scraping. Processamento de Linguagem Natural.
9	07/05	(E2): Etapa 2 do Projeto Final de Semestre. Classificação Textual.
10	14/05	Classificação Textual. Validação Cruzada.
11	21/05	Sistemas de Extração da Informação. Sumarização de Documentos.
12	28/05	Análise de Sentimentos. Reconhecimento de Entidades Nomeadas.
13	04/06	(E3): Etapa 3 do Projeto Final de Semestre. Clusterização.
14	11/06	Clusterização.
15	18/06	Similaridade de Documentos. Métricas de Similaridade.
16	25/06	Retirada de dúvidas sobre a Etapa Final do projeto.
17	02/07	(EF): Etapa Final do Projeto Final de Semestre.
18	09/07	Prova de Recuperação.

**Cronograma sujeito a alterações e adaptações ao longo do semestre.*