



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



PLANO DE ENSINO

1. IDENTIFICAÇÃO
Disciplina: CIN7502 - Mineração de Texto Carga Horária: 36 H/A - 2 créditos Oferta: Obrigatória para o Curso de Graduação em Ciência da Informação. Horário: Segundas-feiras, das 18h30 às 20h10 Local: CCE-CCE131 Prof. Ivam Galvão Filho - ivam.galvao.ufsc@ufsc.br Atendimento extraclasse: Segundas-feiras, das 14h – 17h na sala 207 bloco C do CED
EMENTA
Técnicas de mineração de texto, Mineração de conteúdo Web, Pré-processamento: stopwords, radicalização, substantivação, NER Processamento: categorização, agrupamento, sumarização, extração de informação, Pós-processamento: métricas de avaliação
2. OBJETIVOS
2.1 Objetivo Geral Capacitar os estudantes a extrair informações relevantes de textos, aplicando técnicas de mineração de dados textuais.
2.2 Objetivos Específicos 2.2.1 Compreender técnicas de mineração de texto; 2.2.2 Preparar dados textuais para análise; 2.2.3 Aplicar técnicas de mineração a diferentes conjuntos de dados; 2.2.4 Avaliar os resultados obtidos; 2.2.5 Executar mineração de texto em diferentes cenários.
3. CONTEÚDO PROGRAMÁTICO
3.1 Introdução à Mineração de Texto 3.1.1 Definição e evolução 3.1.2 Tipos de dados 3.1.3 Padrões de representação 3.1.4 KDT (Knowledge Discovery from Text) 3.2 Desenvolvimento da Mineração 3.2.1 Preparação dos dados 3.2.1.1 Análise léxica 3.2.1.2 Bag of Words 3.2.1.3 Modelo de espaço vetorial 3.2.1.4 TF-IDF 3.2.2 Recall e precisão 3.3 Mineração de Dados Textuais na Web 3.3.1 Web crawlers 3.3.2 PLN 3.3.3 Análise de redes sociais 3.3.4 Mineração de opiniões e sentimentos 3.4 Machine Learning para Textos 3.4.1 Aprendizado supervisionado e não supervisionado 3.4.2 Classificação de textos 3.4.3 Ordenamento por relevância
4. BIBLIOGRAFIAS
4.1 Bibliografia Básica CHATTERJEE, Rohit. Top 7 Free NLP Books To Read. Disponível em: https://analyticsindiamag.com/top-7-free-nlp-books-to-read/ . Acesso em: 05 ago. 2020. DOWNEY, Allen. Think Python: How to Think Like a Computer Scientist. Needham, MA: Green Tea Press, 2012. Disponível em: http://greenteapress.com/thinkpython/thinkpython.pdf . Acesso em: 05 ago. 2020. FELDMAN, Ronen; SANGER, James. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge, MA: University Press, 2007. 410 p. Disponível em: http://wtlab1.um.ac.ir/images/e-library/text_mining/The%20Text%20Mining%20HandBook.pdf . Acesso em: 05 ago. 2020.



MICROSOFT DEVELOPER. Python for Beginners, 2020. Disponível em: <https://www.youtube.com/playlist?list=PLlrXD0HtieHhS8VzuMCfQD4uJ9yne1mE6>. Acesso em: 05 ago. 2020.

PYTHON.ORG. O Tutorial Python, 2020. Disponível em: <https://docs.python.org/pt-br/3/tutorial/index.html>. Acesso em: 05 ago. 2020.

SCIFORCE SOLUTIONS. Top 10 Books on NLP and Text Analysis, 2020. Disponível em: <https://www.kdnuggets.com/2019/01/top-10-books-nlp-text-analysis.html>. Acesso em: 05 ago. 2020.

THOMAS, Stephen. Text Mining and Natural Language Processing Resources, 2020. Disponível em: https://github.com/stepthom/text_mining_resources. Acesso em: 05 ago. 2020.

W3SCHOOLS. Python Tutorial, 2020. Disponível em: <https://www.w3schools.com/python/>. Acesso em: 05 ago. 2020.

4.2 Bibliografia Complementar

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

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.

FINLAY, S. *Predictive Analytics, Data Mining, and Big Data: Myths, Misconceptions and Methods*. Hampshire (UK): Palgrave Macmillan, 2014. 248 p.

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.

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*. 3rd ed. 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.

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.

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

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

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

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

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



5. METODOLOGIA

- As aulas serão ministradas no formato presencial e incluirão também atividades extra-classe de forma assíncrona a serem publicados no Moodle.
- Teremos como apoio da disciplina o Moodle institucional da UFSC (<http://moodle.ufsc.br>) e o mesmo será utilizado como instrumento de suporte da disciplina, devendo portanto, ser acessado todas as semanas pelos alunos.
- A organização dos conteúdos se dará através de tópicos, sendo que cada um terá um fórum de discussão (utilizando a funcionalidade já disponível no AVA Moodle) para que os alunos possam debater suas opiniões e dúvidas.
- Aulas expositivas, com discussão em classe dos tópicos apresentados e exercícios práticos e/ou escritos

6. AVALIAÇÃO

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

- A frequência mínima obrigatória é de 75% das aulas.
 - O aluno deverá entrar no Moodle e marcar sua presença.
- Em caso de falta em dia de avaliação (prova, apresentação de trabalho ou mini-seminário), encaminhar justificativa e pedido formal à Chefia do Departamento de Ciência da Informação, no prazo de três (3) dias úteis.
- O aluno que obtiver nota final entre 3,0 e 5,5 e frequência suficiente poderá, ao final do semestre, realizar uma prova de recuperação de todo o conteúdo.
- A nota final será calculada a partir da média simples entre a nota final obtida durante o semestre e a nota obtida na prova de recuperação.

Avaliações:

- Serão realizados diversos exercícios durante o semestre (E1).
- Será realizada uma avaliação teórica (A1).
- Será realizado um trabalho final (T1).
- A média final (MF) será calculada conforme a seguinte fórmula:
$$MF = (E1 + A1 + T1) / 3$$
- E1: Todos os exercícios devem ser entregues via Moodle. Haverá de 1 à 3 listas e as entregas poderão ser parciais em cada atividade extra-classe.
- A1: A avaliação será presencial e entregue até o fim da aula.
- T1: Um trabalho escrito entregue via Moodle e uma apresentação do

conteúdo.

OBSERVAÇÃO:

* Frequência: Obrigatório no mínimo em 75% das aulas (resolução 017/Cun de 30/09/97). * No caso de falta no dia programado para a entrega das avaliações, encaminhar justificativa e pedido formal à chefia do Departamento, no prazo de até três dias úteis, conforme a resolução nº 17/CUN/97, de 30/09/97.

7. CRONOGRAMA

Aula	Data	Conteúdo	Leitura/Atividade
1.	11/08	Apresentação do plano de ensino	Apresentação com slides da disciplina e do plano de ensino
2.	18/08	Introdução à Mineração de Texto	Apresentação expositiva com slides e exercícios de fixação
3.	25/08	Desenvolvimento da Mineração I	Apresentação expositiva com slides e exercícios de fixação
4.	01/09	Atividade prática de exercícios	Lista de exercícios
5.	08/09	Desenvolvimento da Mineração II	Apresentação expositiva com slides e exercícios de fixação
6.	15/09	Atividade prática de exercícios	Lista de exercícios
7.	22/09	Mineração de Dados Textuais na Web I	Apresentação expositiva com slides e exercícios de fixação
8.	29/09	Atividade prática de exercícios	Lista de exercícios
9.	06/10	Mineração de Dados Textuais na Web II	Apresentação expositiva com slides e exercícios de fixação
10.	11/10	Atividade extra-classe	Leitura de artigos, elaboração de trabalho e entrega de lista de exercícios
11.	13/10	Atividade prática de exercícios	Lista de exercícios
12.	20/10	Machine Learning para Textos	Apresentação expositiva com slides e exercícios de fixação
13.	03/11	Elaboração de trabalho e tutoria	Elaboração de trabalho T1
14.	10/11	Apresentação do trabalho – (T1)	Apresentação de trabalho prático
15.	17/11	Apresentação do trabalho – (T1)	Apresentação de trabalho prático



16.	24/11	Revisão	Apresentação expositiva com slides e exercícios de revisão
17.	01/12	Avaliação (A1)	Avaliação sobre todo o conteúdo da disciplina
18.	08/12	Recuperação final	Avaliação de recuperação sobre todo o conteúdo da disciplina