Information Retrieval and Text Mining

Nuno Escudeiro (nfe@isep.ipp.pt)

Ricardo Almeida (ral@isep.ipp.pt)

Planning

Planning

S	T (2 HORAS/SEMANA)	TP (1 HORA/SEMANA)	PL (2 HORAS/SEMANA)
1 24 – fev	- Introdução à Recuperação de Informação (IR), Filtragem de Informação (IF), Extração de Informação (IE) e Mineração de Texto - Processo geral da recuperação de informação. - Indexação e consulta/pesquisa de informação não estruturada	Componentes de um sistema de recuperação de informação	Lançamento Projeto e Seminário. Criação dos grupos. Seleção de temas. Preparação do artigo sobre os temas selecionados
2 3 - mar	- Avaliação em recuperação e extração de informação. Precision, recall, F1-score - ROC curve, AUC, Precision/Recall curve	Avaliação em recuperação e extração de informação	Preparação do artigo
3 10 - mar	- Boolean Retrieval Model: Basics of Boolean algebra and operations in information retrieval - Conjunctive and Disjunctive Queries: Understanding AND, OR, and NOT operators	Apresentação de informação, visualização	R R packages for IR R packages for TM
4 17 - mar	- Técnicas de indexação: Inverted index creation and compression - Posting lists: Representação e manipulação de posting lists	Construção de índices	Lucene
5 24 - mar	- Retrieval Models: Probabilistic and language modeling approaches - Ranking Algorithms: Okapi BM25 and Language Model for Information Retrieval (LMIR) - Relevance Feedback: Techniques for improving search results based on user feedback. - Query Expansion: Expanding queries to enhance retrieval effectiveness	Modelação de textos e conjuntos de textos (corpora)	Métricas de avaliação: Precision, recall, F1-score, and their significance Evaluation Metrics: ROC curve, AUC, Precision/Recall curve Entrega dos artigos - IR Seminar

Planning

S	T (2 HORAS/SEMANA)	TP (1 HORA/SEMANA)	PL (2 HORAS/SEMANA)
1 24 – fev	- Introdução à Recuperação de Informação (IR), Filtragem de Informação (IF), Extração de Informação (IE) e Mineração de Texto - Processo geral da recuperação de informação. - Indexação e consulta/pesquisa de informação não estruturada	Componentes de um sistema de recuperação de informação	**Lançamento Projeto e Seminário. Criação dos grupos. Seleção de temas. Preparação do artigo sobre os temas selecionados**
2 3 - mar	- Avaliação em recuperação e extração de informação. Precision, recall, F1-score - ROC curve, AUC, Precision/Recall curve	Avaliação em recuperação e extração de informação	Preparação do artigo
3 10 - mar	- Boolean Retrieval Model: Basics of Boolean algebra and operations in information retrieval - Conjunctive and Disjunctive Queries: Understanding AND, OR, and NOT operators	Apresentação de informação, visualização	R R packages for IR R packages for TM
4 17 - mar	- Técnicas de indexação: Inverted index creation and compression - Posting lists: Representação e manipulação de posting lists	Construção de índices	Lucene
5 24 - mar	- Retrieval Models: Probabilistic and language modeling approaches - Ranking Algorithms: Okapi BM25 and Language Model for Information Retrieval (LMIR) - Relevance Feedback: Techniques for improving search results based on user feedback. - Query Expansion: Expanding queries to enhance retrieval effectiveness	Modelação de textos e conjuntos de textos (corpora)	Métricas de avaliação: Precision, recall, F1-score, and their significance Evaluation Metrics: ROC curve, AUC, Precision/Recall curve **Entrega dos artigos - IR Seminar**

Planning

S	T (2 HORAS/SEMANA)	TP (1 HORA/SEMANA)	PL (2 HORAS/SEMANA)
6 31 - mar	Apresentação dos temas escolhidos pelos grupos para o sistema de IR/TM a implementar.	Avaliação dos artigos - IR Seminar	Avaliação dos artigos - IR Seminar Iniciar fase 2 (projeto IR/TM).
7 7 - abr	- Processo de recuperação de informação na web, motores de pesquisa -Algoritmo PageRank; Algoritmos: Introdução à análise de hiperlinks e grafos web. - Recuperação de informação centrada no utilizador - Sistemas de recomendação Collaborative filtering	Modelação de textos e conjuntos de textos (corpora). Avaliação de sistemas de recuperação de informação. Avaliação de Sistemas de Recomendação: Métricas para avaliar a qualidade das recomendações	Projeto IR/TM
8 14- abr	- Processo de recuperação de informação na web, motores de pesquisa -Algoritmo PageRank; Algoritmos: Introdução à análise de hiperlinks e grafos web. - Recuperação de informação centrada no utilizador - Sistemas de recomendação Collaborative filtering Páscoa (MINTRI tem aulas à 6ªf)	Modelação de textos e conjuntos de textos (corpora). Avaliação de sistemas de recuperação de informação. Avaliação de Sistemas de Recomendação: Métricas para avaliar a qualidade das recomendações Páscoa (MINTRI tem aulas à 6ªf)	Páscoa (MINTRI tem aulas à 6ªf)
21-abr	Páscoa	Páscoa	Páscoa
9 28 - abr	- Modelação de dados semiestruturados e não estruturados. Modelação de textos e conjuntos de textos (corpora). Dicionários, Estruturas de dados, Tolerância de termos - Term Frequency-Inverse Document Frequency (TF-IDF) - Vector Space Model	Medidas de similaridade entre documentos e consultas	Caso de estudo de mineração de texto

Planning

S	T (2 HORAS/SEMANA)	TP (1 HORA/SEMANA)	PL (2 HORAS/SEMANA)
6 31 - mar	**Apresentação dos temas escolhidos pelos grupos para o sistema de IR/TM a implementar.**	**Avaliação dos artigos - IR Seminar**	**Avaliação dos artigos - IR Seminar** **Iniciar fase 2 (projeto IR/TM).**
7 7 - abr	- Processo de recuperação de informação na web, motores de pesquisa -Algoritmo PageRank; Algoritmos: Introdução à análise de hiperlinks e grafos web. - Recuperação de informação centrada no utilizador - Sistemas de recomendação Collaborative filtering	Modelação de textos e conjuntos de textos (corpora). Avaliação de sistemas de recuperação de informação. Avaliação de Sistemas de Recomendação: Métricas para avaliar a qualidade das recomendações	Projeto IR/TM
8 14- abr	- Processo de recuperação de informação na web, motores de pesquisa -Algoritmo PageRank; Algoritmos: Introdução à análise de hiperlinks e grafos web. - Recuperação de informação centrada no utilizador - Sistemas de recomendação Collaborative filtering Páscoa (MINTRI tem aulas à 6ªf)	Modelação de textos e conjuntos de textos (corpora). Avaliação de sistemas de recuperação de informação. Avaliação de Sistemas de Recomendação: Métricas para avaliar a qualidade das recomendações Páscoa (MINTRI tem aulas à 6ªf)	Páscoa (MINTRI tem aulas à 6ªf)
21-abr	Páscoa	Páscoa	Páscoa
9 28 - abr	- Modelação de dados semiestruturados e não estruturados. Modelação de textos e conjuntos de textos (corpora). Dicionários, Estruturas de dados, Tolerância de termos - Term Frequency-Inverse Document Frequency (TF-IDF) - Vector Space Model	Medidas de similaridade entre documentos e consultas	Caso de estudo de mineração de texto

Planning

S	T (2 HORAS/SEMANA)	TP (1 HORA/SEMANA)	PL (2 HORAS/SEMANA)
5 - mai	Queima das Fitas	Queima das Fitas	Queima das Fitas
10 12 -mai	- Pré-processamento de texto (stop words, stemming, seleção de atributos, thesauri, etc) - Pré-processamento de texto (stop words, stemming, seleção de atributos, thesauri, etc)	Ferramentas, linguagens e bibliotecas (R, Wordnet, , etc)	Projeto IR/TM
11 19 -mai	- Operações sobre texto (POS tagging, NER, ...) - Operações sobre texto (POS tagging, NER, etc)	Ferramentas, linguagens e bibliotecas (R, Wordnet, etc)	Projeto IR/TM
12 26 -mai	- Aprendizagem automática na recuperação de informação - Aplicações (classificação, clustering, sentiment analysis, topic drift, etc) - Classificação de texto	Tendências na área da recuperação de informação e mineração de texto Considerações éticas	Projeto IR/TM
13 2 - jun	Projeto IR/TM	Projeto IR/TM	Projeto IR/TM.
14 9 - jun	**Entrega do projeto.** **Avaliação**	**Avaliação**	**Avaliação**
15 16 -Jun	**Avaliação**	**Avaliação**	**Avaliação**
16 23 -Jun	**Avaliação**	**Avaliação**	**Avaliação**

Evaluation

Brief description, purpose

Evaluation

Components	Type	Weight	Min	Repeatable 1st Exam	Repeatable 2nd Exam
M1 – Seminar	Paper	25	9,5 (weighted average)	NR	NR
M2 – Project	Project	45		NR	NR
EG – Exam	Written	30	8.0	Yes	Yes

- Final Classification (CF)
  - CF=M1*0,25 + M2*0,45 + EG *0,3
- Partial Improvement Repeatable Component (MP) - EXAM
  - CF=MNR*0,7 + MP *0,3