startfactor
statistical consulting and training

Text Mining

www.startfactor.pt

Descrição do serviço

A análise de texto – *text mining* – é, hoje em dia, consensualmente reconhecida como uma poderosa ferramenta analítica de elevado valor no mundo do negócio.

Apesar disso os dados de texto, incluindo emails, documentos, jornais ou revistas, conteúdo de redes sociais (*unstructured data*) são tipicamente arquivados sem serem analisados desperdiçando-se assim a oportunidade de produzir informação significativa em termos do seu conteúdo.

O reconhecimento de padrões embutidos nos *terabytes* de dados não estruturados começa, em regra, com a pré-exploração do material com vista à seleção do *corpus* a ser analisado. Numa primeira fase pretende-se o contacto geral com os documentos em análise, visando a compreensão do contexto, conceitos e significados fundamentais. Numa segunda fase é feita a seleção das unidades de análise (ou unidades de significados). Aqui serão tidas em conta as questões de investigação específicas bem como as estruturas teóricas de modelos organizacionais. Segue-se a fase de categorização e sub-categorização (*Text Categorization*), baseada em critérios frequencistas (categorização não supervisionada) e de relevância contextual (categorização supervisionada). Nas fases seguintes procede-se à construção da base de dados que permitirá a aplicação de métodos quantitativos de análise da informação recolhida documentalmente.

Uma grande quantidade de dados é importante, mas fundamental é saber como tratar os dados para optimizar o uso da informação neles contida. O tratamento da informação extraída das fontes documentais é tipicamente realizada por técnicas de *data mining*. Em regra, o estudo tem início pela análise exploratória da matriz de termos por métodos de contagem e quantificação de associações (*count-based evaluation*). A análise é feita por métodos numéricos e métodos gráficos (e.g., correlation plots e word clouds). As técnicas de análise de clusters usadas em conjunção com técnicas de análise factorial permitem agrupar e classificar a informação.

O estudo de tendências temporais no uso de palavras, expressões e/ou frases e, caso haja interesse, a análise de citação de entidades/individualidades (e.g., nomes, lugares, datas, preços, etc.) possui, em geral, grande interesse organizacional.

O uso de técnicas de *Predictive Data Mining* permite completar a análise com a modelação de uma variável resposta (de interesse), tendo em conta associações significativas encontradas com base na informação documental.

A análise deste tipo de informação tem múltiplas aplicações indo desde a simples e tradicional deteção de email spam e categorização de mensagens de email até aplicações complexas que incluem o estudo de fraudes literárias, monitorização de redes sociais, deteção de novas oportunidades de venda e investimentos e análise de comportamento do consumidor, através da identificação de arquétipos.