Text Mining: o uso de dados não-estruturados como insumo para extração de inteligência

Text Mining: the use of unstructured data as input for intelligence extraction

Moisés Lima Dutra[a]

Transcrição da apresentação

Bom dia, nós vamos conversar hoje sobre Text Mining. Essa é uma disciplina que eu ministro lá na Universidade Federal de Santa Catarina no curso de Ciência da Informação, então tentei fazer um apanhado de maneira geral o tema, um campo muito amplo e vasto então na verdade não teríamos tempo de entrar nos detalhes, de cada um dos processos e de cada uma das etapas que ocorrem em um Text Mining.

Eu vou tentar fazer de maneira geral um overview e a ideia aqui é que vocês saiam sabendo do que se trata, e aí a gente pode depois desenvolver isso mais pra frente pra quem tiver interesse a gente pode conversar alguma coisa e tentar aprofundar ou detalhar alguma coisa que tenha passado.

Essa aqui é a agenda da apresentação, eu começar primeira uma introdução, um preâmbulo, depois a gente vai dar uma olhada geral no que é o Text Mining, vou apresentar para vocês as etapas de mineração, alguns exemplos e falar no final sobre desafios e oportunidades.

Vamos começar falando de cinema, vocês gostam de cinema? Já viram esse filme, Contato de Risco? a princípio tem tudo para ser um grande filme, lançado em 2003, título original Gili, título no Brasil ficou Contato de Risco, gênero comédia romântica, vejam só que o elenco é um grande elenco, Ben Affleck, Jennifer Lopez, Christopher Walken, Al Pacino. Por que esse filme é tão emblemático? Ele acabou criando um antes e depois, muito simples.

Porque ele custou 75 milhões de dólares e arrecadou 7,2 milhões de dólares, um grande prejuízo apesar de todo aquele elenco. Então isso aqui acendeu uma luz vermelha em Hollywood, o filme ganhou 6 prêmios no Grammy de Ouro: pior filme, pior diretor, pior ator, pior atriz, pior roteiro, etc. Aí além disse foi nomeado em algumas outras categorias: pior ator e atriz coadjuvante, e para finalizar ganhou o prêmio de pior comédia dos últimos 25 anos do Grammy.

O filme foi de fato um marco. Então isso acende uma luz amarela nos produtores de Hollywood, nos criadores de roteiro no sentido será que a gente não consegue prever o sucesso do filme? Será que a gente tem como saber se o roteiro é certo não, se seria adequado ou não? Além disso, será que eu poderia estar extrair então a informação, extrair inteligência de maneira a não jogar dinheiro numa obra que não vai dar retorno? Por exemplo, os orçamentos podem ser baseado nas possibilidades de sucesso? O que uma obra, enfim, pre diz.

E aí é interessante falar do caso Epagogix. A Epagogix é uma empresa britânica, que foi fundada em 2003, ela utiliza softwares analíticos baseados em redes neurais, e ela faz uma coisa

a Universidade Federal de Santa Catarina (UFSC). E-mail: moises.dutra@ufsc.br. ORCID: https://orcid.org/0000-0003-1000-5553. Currículo: http://lattes.cnpq.br/1973469817655034

muito interessante, ela avalia o grau sucesso dos roteiros e das tramas criadas pelos escritores, pelos criadores de roteiro.

Então ela começou trabalhando com filmes e programas de TV, na Inglaterra, no Reino Unido, logo no início, veja que interessante foi em 2003 que foi o mesmo ano de lançamento do filme. São dois eventos que se conectam mas que a princípio não tinham nenhuma conexão.

Então a Epagomix estabeleceu uma parceria com órgãos próprios da Europa no sentido de avaliar 16 programas piloto de televisão, eles queriam levar os programas para o mercado americano e como a gente sabe o público americano tem outro perfil. Portanto o roteiro que vai para os Estados Unidos ele precisa ser diferente do roteiro que é feito para Europa.

Vejam que interessante, para 13 programas a taxa de audiência calculada ficou dentro da margem de erro de dois pontos percentuais das taxas de audiências reais medidas quando o programa foi ao ar. E desses 13 programas, 6 programas ficaram dentro da margem de erro e 0,06.

Os caras pensaram, pera aí, a gente tem aqui uma coisa muito interessante, a gente consegue a partir da análise do script, identificar a quantidade de audiência do programa e, obviamente, isso vira um modelo de negócio onde você pode direcionar orçamento e procurar investidores.

Na verdade a Epagomix foi a primeira de uma série de empresas e ela foi o estopim daquilo que se costumou chamar em inglês de Análise de Scripts. Hoje existem diversas opções e uma das mais conhecidas é o Script Book que cobra 100 dólares por cada piloto que você faz de um roteiro inglês.

Cada um de nós que podemos escrever um roteiro tentando ganhar um dinheiro em Hollywood faz o upload no Script Book para tentar avaliar este roteiro. Então eu trouxe dois screenshots do dashboard do Script Book, onde são feitas análises do tipo do filme, do gênero, da categorização do filme como a variedade a partir de critérios norte-americanos, a estimativa de audiência mais ou menos, ela faz tipo uma tag cloud aqui na sinopse do filme.

Além disso são feitas algumas avaliações dos personagens com relação a diversas facetas, as diversas características como raiva, tédio, felicidade, amor, e aí sim você consegue avaliar pelos scripts a quantidade de cada uma dessas características presentes no personagem da sinopse.

No Script Book é possível analisar a probabilidade de sucesso de um filme, por meio dessa análise de sentimentos e dessa análise da história. Como é que os caras fizeram isso?

Tem um modelo de Machine Learning, eles pegaram um dataset composto por todos os filmes exibidos em cinema nos EUA desde 1970, foi por puro reconhecimento de padrão, se pegou os filmes de sucesso e cruzou com os filmes com sucesso de bilheteria.

Os filmes de maior sucesso foram usados então como um sistema de treinamento, e a partir daí cada novo script que é submetido ao site é comparado com esse sistema de treinamento e é então submetido a um modelo de classificação feito para prever o grau de sucesso que esse filme pode ter.

Como nem tudo são flores, esse é um site bastante criticado principalmente pelos autores que se sentem pouco, digamos, no terraço de uma máquina, os caras têm reviews bastante curiosos sendo interessante a gente procurar, os caras dizendo que a história é a capacidade criativa, mas enfim… Sem entrar no mérito o fato é que isso existe e já está sendo utilizado.

Em publicações científicas, esses dois carinhas aqui eram estudantes do MIT em 2005, no Departamento de Sistemas Cibernéticos e Informática. Mulecada nova, começou a desenvolver programas e eles fizeram alguns artigos científicos, como esse daqui que é muito interessante, propondo uma metodologia. É um artigo bem interessante, bem estruturado, ele trabalha um pouco a questão dos relógios de Lamport, a cadeia de Markov, são alguns os matemáticos.

O que há de tão interessante sobre este artigo? Ele é mentira, é fake.

Os três estudantes descobriram uma maneira de criar um gerador automático de artigos científicos para a Ciência da Computação. Eles não param por aí. Submeteram o artigo a um evento científico e o artigo foi aceito neste evento aqui, e alguns dias antes do evento eles vieram e confessaram tudo por uma Rede Social e disseram "ah pessoal era tudo uma brincadeira, o nosso artigo não é de verdade.

Isso gerou um grande zunzunzum, eles disseram que haviam criado um gerador automático de ciberespaço da Ciência da Computação, então você coloca o nome dos autores, alguma proposta de tópico e a partir daí um software gera automaticamente o paper e a prova de fogo foi justamente porque aquilo passou num sistema de avaliação por pares, de uma conferência bem-conceituada.

A I3E patrocinava o evento e retirou o patrocínio no mesmo ano, a empresa não gostou e começou a analisar então o software que foi a público. Só entre 2008 e 2013 de mais de 120 prêmios criados com software foram detectados pela I3E. Isso começou de uma forma a se tornar tão freqüente que daí que surgiu o software que detectava papers criados assim.

Então a Springer, que é aquela editora, se juntou à universidade de Grenoble, na França, e junto eles criaram um cyber detect, que é um detector de artigos gerados com o software.

Coloquei aqui mais alguns exemplos de geradores de texto, mas existem aos montes, como um clássico gerador de lero-lero né, usados para engrossar TCCs, Teses e Dissertações.

Mas o que está por trás então desses dois processos, desses dois cenários, tanto de cinema quanto da publicação científica? Simples, a detecção de padrões, a análise textual profunda e muito forte procurando buscar padrões e a partir desses padrões aplicar uma análise para extrair uma parte da inteligência, para extrair conhecimento.

Tanto o Script Book quanto a Epagomix fazem análise de roteiros de cinema. O Script Book tem um dataset monstruoso, com todos os filmes americanos. A Epagomix utilizava um outro método, que é preciso abrir um parenteses aqui.

Desde os anos 70 nos Estados Unidos, roteiristas de hollywood sabem que existe uma estrutura nos filmes, então foi escrito um livro chamado A Jornada do Herói, por um estudioso do folclore, enfim, e escreveu esse livro dizendo o seguinte: a humanidade sempre conta a mesma história, todas as nossas histórias são fazendas de uma mesma história, e ele faz um paralelo com personagens da história, Buda, Jesus Cristo, Confúcio, depois ele vem para personagem mais modernos. E os roteiristas de Hollywood entenderam um pouco isso, e o que se diz é que todos os filmes fizeram sucesso nos últimos 40 e 50 anos, eles seguem de alguma forma essa estrutura.

Cada sinopse teria 3 atos. O primeiro ato duraria 25% do filme, com a apresentação dos personagens, o personagem sendo uma pessoa comum e vivendo um dia-a-dia banal.

Na página dois do script é interessante tem um fato, uma revolução que leva para o ato 2, que pode ser uma coisa traumática ou uma coisa excepcional, se for filme de ação e aventura pode ser um momento na estrada que gera então para o herói uma busca por algo que quase espiritual, segundo alguns critérios.

Eu to detalhando todo o ato 2, o antagonista do filme seria uma serial que confrontaria o herói pessoalmente, e teria que superar essa meta se pode chegar um grau superior, e esse ato 2 se encerraria com isso.

Então o nosso herói vencendo seguiria para o ato 3 que seria trazer esses novos aportes que ele absolveu nesta sua jornada para aquela sua vida medíocre lá no início. Então ele voltaria para aquela vida com esses novos aportes. Então esse é a Jornada do Herói, e o autor do Script Book trabalhou em cima disso. Independentemente da metodologia que se usa, o fato é que as sinopses de filme hoje estão sendo avaliados por ferramentas e veio para ficar.

Quanto à análise de produções científicas, a detecção de padrões também está sendo utilizado. Na prática a gente precisa disso porque humanamente é impossível vocês vão avaliar devido ao tamanho e a quantidade de documentos e texto. Basicamente o que a gente tem por trás é o Text Mining permeando o fundamento desses cenários.

Mineração de texto, também chamado por alguns autores de mineração de dados, pessoais, a gente encontra isso também. É uma busca por padrões em textos digitais com linguagem natural, uma definição bem sucinta e objetiva

O texto é de linguagem natural, que é a mais importante fonte de informação, tem um relatório interessante de 2011 que estimava que mais de 90% dos dados existentes hoje são dados textuais, são dados não estruturados. Então vejam agora, vamos pensar um pouco nos nossos sistemas de trabalho com dados estruturados, como banco de dados, os metadados, seria apenas 10%, a gente está na ponta do iceberg.

Então a gente tem ainda 90% de dados que de forma nenhuma estão organizados. Mas, nem por isso, deixam de ser tão ou mais importante quanto os dados que estão estruturados.

Os dados textuais estão em sua maioria na forma não estruturados, uma sintática que não é confiável, justamente porque eu não tenho nenhum tipo de regra como um sistema de informação.

Desde os anos 50 e 70 fazer um tradutor entre o Inglês e o Russo, por exemplo. Porque a gente nunca conseguiu isso 100%? Porque trabalhar com texto é uma coisa bastante complexa, tem diversas nuanças e características que são muito peculiares apesar da grande evolução dos últimos tempos, como o Google Translator que está aí pra não me deixar mentir, mas, ainda assim, a gente precisa ir lá no Translator e fazer ajustes, retirar redundância, e ajustar o contexto.

Text Mining ou Data Mining? Essa é uma dúvida que muita gente possui, ah mas isso aí de minerar texto não é a mesma coisa de minerar dados? Sim e não.

É óbvio que nós temos hoje um ambiente com abundância de dados, só para citar algumas coisas, tudo que é informação digital, ela é um dado obviamente. Então imagine quantas transações eletrônicas são efetuadas por dia e por hora, todos os dados que estão na nuvem, todos os dados que estão em empresas, que estão universidades ou institutos de pesquisa, toda a

parte de internet das coisas e computação pervasiva, é cada vez mais dispositivos e sensores gerando dados, Machine Learning, etc.

É óbvio que a gente está num ambiente de abundância de dados, o pessoal coloca Data Mining nos textos, a gente encontra uma frase que é muito interessante principalmente nos Estados Unidos que dizia o seguinte: mostre-me dados que eu apresentarei padrões.

Mineração de dados ela tem alguns pré-requisitos no entanto, quais são? Os datasets precisam ser muito bem estruturados e organizados. Em toda a fase de preparação de um Data Mining ela é importantíssima, senão os resultados que obtemos não vão ser corretos, imprecisos. Eu preciso ter formatos muito bem definidos, ou seja, o processo prévio de preparação tem peso importante, então ou os dados já estão organizados para serem minerados ou a gente vai ter que fazer isso na mão.

Se a gente pudesse contextualizar as duas definições dos conceitos, apresentaria dessa forma aqui, a gente tem os dados não estruturados sendo estruturados.

O Data Mining, pessoal, ocorre em bancos de dados estruturados, SQL, Oracle, DB2, existe um formato, um padrão formal a ser seguido, esse dado já está organizado.

Mas a gente tem também os dados não estruturados, que é texto em linguagem natural, que é um post no Facebook ou no Twitter, que é a frase no WhatsApp que a gente escreve sem se preocupar com vocabulário controlado, tesauro ou ontologia, a gente escreve errado, com gírias, com pontuação incorreta, enfim…

E a gente tem uma camada cinzenta aqui entre esses dois, que eu chamaria de dados semi-estruturados, que são quase não-estruturados, que seriam por seriam, por exemplo, arquivos em XML de um sistema proprietário.

Hoje, os metadados e as tags foram definidas para funcionar assim. Ele é estruturado, mas é semi-estruturado, porque dependendo do sistema com o qual eu vá processar aquelas dados ele vai entender ou não.

Então Text Mining abrange essas duas camadas, dos dados estruturados e dos dados semi-estruturados.

Mas o que é interessante é o seguinte, na verdade elas interconectam no sentido de que mesmo trabalhando com dados não estruturados, muitas vezes no Text Mining o objetivo é chegar aqui, muitas vezes quando começo analisar texto o que eu quero na verdade é chegar a um modelo de dados estruturados, pode ser que a pura análise, a pura varredura do texto tal como estão, me permita extrair os ensaios que eu busco. Caso contrário, eu vou ter que fazer essa transição, então essa transição do dado não estruturados para os dados estruturados nada mais é que organizar, preparar e harmonizar esses dados, higienizar enfim, existem outras palavras que se usam para que eu cheguei nesses dados aqui.

Se a gente pudesse, digamos, subir uma cama nesse overview, eu posso se contentar com a Ciência de Dados, ela vai abarcar todos os dados, vai abarcar Text Mining, vai abarcar o Data Mining.

A transformação desse texto cru ou texto bruto em números, porque na verdade esse processo aqui ele acaba de certa forma transformando o que é texto em números, porque eu posso não ter chegado a uma estrutura em comum mas eu posso aplicar modelos matemáticos, aí

eu consigo trabalhar algum tipo de Machine Learning, porque aí tem números e eu consigo processar matematicamente, enquanto é só texto eu não consigo.

Na verdade essa transformação do texto cru envolve uma série de outras coisas. Eu posso tirar uma série de coisas, medição, enfim… De uma maneira geral essa seria a idéia dos Text Mining.

Esse processo de transformar o texto em números, a raiz do processo dá-se o nome de vetorização, eu vou vetorizar esse texto, vou transformar isso em um vetor matemático, Isso aqui é um assunto muito vasto, isso aqui a gente trabalha em muitas aulas lá na UFSC mas eu vou tentar pincelar aqui de uma maneira bem sucinta do que significa sem entrar muitos em detalhes técnicos.

É importante falar do modelo SAC of Words, ou modelo SAC de palavras. Esse é o modelo que está por trás da minha mineração de texto. Esse modelo compreende a freqüência das palavras distintas existentes nos documentos. Então na verdade você vai pegar um conjunto de dados no qual cada elemento desse conjunto é a freqüência de uma palavra em determinado documento.

Evidentemente, eu tendo um conjunto de dados nesse para cada um deles. Essa freqüência de termos gera um peso que vai se associado a cada um dos termos. Se conta o número de ocorrências do termo dentro do documento para me gerar um valor matemático que a gente chama de freqüência de Deus, e isso vai atribuir o peso a quantidade de ocorrências desse termo.

Entretanto, existe um problema aqui. O meu cálculo final pode ficar desequilibrado, porque dependendo do contexto que eu estiver trabalhando há termos que são figurinhas carimbadas, que existem aos montes, e isso pode enviesar o resultado final.

Para que isto não ocorra, existe o inverso da frequência dos termos nos documentos, que é um outro cálculo que se faz. Ele tenta melhorar esse problema de que todos termos se torne proeminente, porque pode se tornar um problema crítico, vou dar um exemplo aqui para vocês.

Imaginem que a gente teve trabalhando com processamento de uma massa documental da indústria automobilística por exemplo, a gente vai ter muito frequentemente a ocorrência dos termos carro, automóvel e veículo, olha se eu já sei que esses termos vão aparecer aos montes e eu ainda sim considerar o peso deles em relação aos outros, vou criar um viés ali que vai impactar firmemente no resultado final que eu quero.

A ideia do universo da frequência e não superdimensionar a ocorrência dos termos que irão ocorrer de qualquer forma, muito frequentemente, então para esse fim a gente considera também o número de documentos no qual eu tenho em uma base, e aí a gente tenta atribuir um peso menor para aqueles documentos que estão mais abrangentemente dispersos, de maneira que eu sei que aquele termo vai aparecer de qualquer forma, eu não preciso minerar porque ele é frequente, então a gente atribui um peso menor.

Como o exemplo da indústria automobilística, imagina que eu tenho lá esses quatro temos, essa coluna representa a quantidade de documentos que esses temas aparecem, e a última coluna a frequência atribuída a eles.

Vejam que o mesmo o termo que foi o que menos apareceu é o que vai ter um peso maior. Por que? Pra que isso se equilibre, o TF que é o universo da frequência vai equilibrar o TF que é a frequência do termo.

A gente chega então no cálculo que a gente quer, que é o cálculo do Tf, que é o produto das duas métricas anteriores, então vamos considerar que a freqüência que o tema aparece em um documento juntamente com o ingresso da freqüência em que ele aparece no documento.

Nesse cálculo aqui ele pretende estabelecer uma relação entre os termos mais freqüentes do documento tanto quanto a relação do documento com o corpus textual como todo.

Ainda assim é preciso que se faça um último ajuste no Tf e ITf, porque eu preciso considerar o padrão de diferentes documentos do mesmo corpus, então imagine que eu tenho um corpus de 10 documentos, e um dos documentos tem dez páginas, e eu tenho outro documento com 100 mil páginas, fica completamente irreal eu atribuir o mesmo peso a esse documento, é óbvio que o documento de 100 mil páginas o termo vai aparecer muito mais frequentemente, a chance dele estourar ali o delta é grande, então a aplica equações matemáticas como aqui mas a gente chega depois nesse ITf que deve ser normalizado. Aí sim esse cálculo que a gente vai associar aos termos dentro do processo de mineração de texto. Aquela vetorização vai ser preenchida para cada termo, geralmente com o cálculo da ITf normalizado.

Algumas aplicações aqui do Texto Mining. Elas são inúmeras, tentei pegar o que era mais significativo aqui, mas isso não é, de maneira alguma, de maneira exaustiva.

Classificação dos documentos, categorização de documentos, sumarização, similaridade, reconhecimento de entidades nomeadas convencionados, análise de sentimentos ou como geração de opiniões, sistemas de pergunta e resposta, chat box, vou tentar rapidamente solucionar cada um deles para que a gente tenha uma ideia do que significa cada uma dessas aplicações.

A classificação de documentos é aplicada quando eu tenho um corpus documental e eu quero literalmente revelar quais os documentos dentro de categorias pré estabelecidas, pré definidos, então o sistema de classificação vai obter sucesso porém ele consegue associar cada termo todo documento a sua correta categoria que corresponde.

Aqui 3 categorias criadas, uma categoria polícia ou pelo telefone celular por filmes né eu tenho documentos que remetem a ao iPhone 6, outro liga da justiça, outra a milk shake de banana, enfim… E a gente roda isso no sistema de classificação textual e ele automaticamente agrupa os textos dentro da categoria pertinente, como é que isso é feito?

Atribui-se um peso para cada um dos documento, avalia-se na hora em que vai definir você faz esse ajuste, a gente pode, por exemplo, ter um exemplo imagine que eu livro que possui mais de 30% do conteúdo sobre a preparação de comida, a gente pode definir se passou dessa categoria, a gente pode dizer o que ele pode se classificar com um livro de culinária de receitas.

Prioridades e pesos são atribuídos. O que é importante lembrar que a classificação é que os grupos eles são criados pelos humanos, eles são pré definidos e aí a gente tem também a uma aplicação similar que é a clusterização.

Na clusterização os grupos são defendidos automaticamente, o próprio processo de mineração textual se encarrega de criar os grupos, então a gente vai aplicar a clusterização sobre

uma massa documental e vai identificar os clusters existentes ali, a gente vai categorizar. Tem um processo a mais, a gente poderia dizer que a classificação é um processo de machine learning supervisionado e aqui é um processo de machine learning não supervisionado.

O objetivo é criar clusters desses documentos que sejam similares, ele procura automaticamente e atribui subgrupos. Mesmo que a granularidade, que estou chamando aqui de documento pode variar desde uma pequena frase até coleções inteiras de documentos.

Aqui é uma ilustração que mostra uma coleção de documentos depois de rodar por clusterizadores e simplesmente criar clusters como agrupamentos desses documentos.

Quais são as vantagens da clusterização? Ela retirou do elemento humano a obrigatoriedade participar desse processo. Muitas vezes você pode por curiosidade ou por identificar tendências também, é muito interessante, você pode ter uma massa documental e quando faltou conversa que eu tô falando também de dados textuais coletados na Web, imagine eu posso estar analisando posts, ou site de notícias, sites informativos, eu quero identificar a tendência, quais são os tópicos mais discutidos naquele momento?

A gente vai rodar aquele de software de clusterização nesse corpus e vai identificar automaticamente os clusters que existentes ali. Principalmente quando o corpus textual tem grande clusteridade, poder dois exemplos aqui na classificação de notícias de última hora. pra ver como funciona com uma notícia trend topics de hoje, documentos no mercado de ações que mudam a cada minuto.

Outra aplicação é a sumarização, ou seja, a criação de resumos automáticos, exatamente isso que vocês estão pensando, imagine que eu tenho um documento e quero criar automaticamente o resumo contendo as partes mais significativas do documento. Ele pode ser aplicado tanto documentos solitariamente quanto em grandes corpos documentais.

A gente consegue definir qual vai ser o tamanho desse resumo, é 10% do total? Claro, é um processo automático, o software precisa saber. Então eu tenho lá um documento de 500 páginas, 50 páginas resume? Será que 50 páginas consegue fazer resumos? Ou 58? 20 páginas?

Além disso a sumarização pode ser feita por quem escolheu os documentos, eu posso simplesmente pegar um documento que seja representativo no meu corpus, então após ter o corpus documental é o momento de dizer, olha esse é um exemplo clássico é indicativa, ele contém resumidamente tudo aquilo que o documento representa.

Não esqueçam de que já foi criado o clusters com o agrupamento de similaridades de documentos, uma outra proposta é sumarizar pela mescla de documentos. Aqui eu tenho três clusters de documentos, vamos pegar o mais significativo de cada um deles e vou criar um novo clusters aqui com o resumo, vou ter um conjunto de documentos que representam de maneira resumida a informação que está no meu cluster.

Algumas vantagens, a gente pode visualizar a sumarização automática de um texto bastante extenso desde que para pesquisas científicas ou uma outra coisa interessante, sumarização automática independente, fazendo ali um fichamento eu posso fazer algo automatizado. Múltiplos documentos com o mesmo tópico, a gente pode pegar de fato aquele que é mais significativo, não ficar repetindo, ou posso organizar automaticamente.

Outra aplicação do Text Mining é a similaridade de documentos. Similaridade de documentos faz uma comparação e indica o grau de similaridade que o documento possui com outras. Então aquele documento 1 e documento 2 eu vou montar a similaridade em cima do escore aqui.

Então a similaridade é realizada como uma etapa anterior a clusterização, você só vai clusterizar documentos similares. O que é mais comum, no entanto, é que isso seja feito como um chamado de documentos compostos, que são aqueles documentos que já vem de vários clusters, de um ou mais clusters, e a gente pode calcular a similaridade desses dois.

Imagine que você tem um corpus documental com 50 mil documentos. Cada documento com n páginas e eu quero clusterizar isso. A gente precisa de alguma forma rodar um algoritmo ali dentro para ele aplicar a similaridade par a par, é uma tarefa bastante custosa na computação, a clusterização é sempre mais pesada processado do que a classificação.

Como é que a gente calcula essa similaridade? Usando nosso retorno atras, a gente chegou aqui na parte da data mining, transformou um texto em números, com números eu consigo calcular a proximidade numérica. Quando transforma o texto em número, eu consigo aplicar processos matemáticos naquilo ali. E aí faltou o Tf e ITf deve ficar com um aqui.

Então geralmente se calcular três valores, eu vejo o valor dos documentos, dos dois documentos mais próximos entre si e o valor dos outros documentos mais distantes entre si, e a média desses valores. A partir daí a gente calcula o score de similaridade.

Vejam aqui um exemplo. Imagina que eu tenho ali dois vetores compostos, só que são três documentos, esse aqui é o único, e a gente vai computar a singularidade daquilo ali e vai chegar nesse escores aqui, 0,50 para o mais próximo, 0,45 o mais distante da média, e 0 o restante, a gente pode dizer de cara de eu tenho um grau médio de 47% de similaridade dos documentos, por exemplo.

A partir dos valores gerados pela similaridade, simplesmente depois faz uma transferência dos documentos para os clusters, a gente agrupá los por similaridade. O cálculo de similaridade pode se bastar por ele mesmo pode ser usado intermediário da clusterização. Existem diversas métricas de similaridades, citei algumas aqui que calcula a distância entre dois strings parecidos, métrica de Manhattan e a distância do seno e coseno, e a distância de Levenshtein.

A distância de Levenshtein ela calcula o número mínimo de edições necessárias em forma de adoções, remoções e substituições. Então a gente observa essas duas strings para saber qual é a distância de uma para a outra.

Outro exemplo aqui são essas palavras, e a distância de Levenshtein é três, porque eu preciso fazer uma substituição aqui, depois uma segunda substituição e finalmente com a inserção de letras, aí eu consigo chegar perto dele.

Então só métricas que são utilizadas no cálculo da similaridade.

As entidades nomeadas são análises que se fazem em textos que a gente consegue extrair daí categorias pré-definidas, como pessoas, organizações, locais, como dados monetários, etc.

Aqui tem um exemplo, imagine que eu rodei essa frase aqui "a Wikipedia corporation é empresa criadora da Wikipedia", e a gente traz essas duas entidades, uma organização que é a Wikipedia Corporation e o software que é a Wikipedia. Vejam aquela frase ali, a gente tem quatro

entidades reconhecidas, então Obama como uma pessoa, a GM como uma organização, Detroit como local e Today como uma data.

Então a gente consegue a partir da extração das entidades nomeadas, veja que eu já estou agregando valor semântico aqui na mineração, eu estou extraindo informação de mais alto valor agregado.

Análise de sentimentos é uma aplicação também bastante utilizada, principalmente quando a gente analisa redes sociais. Basicamente é uma extração subjetiva da informação, a gente tem duas formas de aplicá-lo, uma forma binária sim ou não, gostou ou não, e uma forma mais elaborada de identificar.

Óbvio que quanto mais especificada é a mineração que eu faço, mais custoso computacionalmente será isso.

Sistemas de perguntas e respostas. Como exemplo o IBM Watson, é uma aplicação que foi feita lá em São Paulo, e conversa com as pessoas que visitam a exposição e ele vai aprendendo também aí. Então existe um sistema de retroalimentação, ele vai reorganizar os documentos baseados em critérios racionais. Muitas vezes pode ser um comentário durante uma resposta, enfim, é um sistema de aprendizagem supervisionada.

E o Chatbox, que são aplicações de conversa entre duas pessoas, eles são baseados no teste de Turing, que foi proposto pelo Alan Turing ainda nos anos 30 para tentar descobrir se uma pessoa conseguiria identificar quando é uma máquina ou uma outra pessoa que estivesse conversando com ela normalmente. Hoje é muito utilizado para serviço de atendimento ao cliente, à venda, no Facebook e ele também vai aprendendo conforme as novas respostas.

Etapas da mineração. A gente tem uma coleta do corpus textual, tem um pré-processamento que é chamado também de processamento de linguagem natural, o processamento em si, e também uma etapa de pós-processamento.

A coleta de corpus textual eu posso trabalhar com documentos que já são meus, ou eu posso acessar datasets que sejam públicos e estejam disponíveis, ou eu posso coletar dados da Web utilizando uma técnica chamada web scraping, que é ir lá e baixar os dados do HTML.

Precisamos ver quais são os documentos relevantes, em que formatos eles se encontram, isso está em ASC, PDF, RDF, etc. Geralmente se transforma para UTF-8 que é padrão para mineração de texto, padrão UNICODE, e aí a gente entra na etapa mais importante que seria o processamento de linguagem natural, que perpassa todas essas etapas aqui.

É uma limpeza textual, uniformização de maiúscula e minúsculas, tokenização textual, expansões e contrações, remoção de símbolos e pontuação, remoção dos stopwords, exposição da grafia, e lematização/radicalização.

A limpeza textual a gente retira tudo aquilo que não serve, tudo aquilo que é estranho ao texto, tokens, anotações, depois a gente uniformiza, o software que compara os strings, geralmente jogam a grafia para o maiúsculo.

Remoção de todos os símbolos de todas as pontuações, afinal de contas isso não agrega nenhum valor semântico.

A Tokenização textual. Os tokens são a unidade básica do Text Mining, são as unidades mínimas e eventualmente serão palavras, mas nem sempre será uma palavra. Eu posso ter uma frase ali, depois que a gente faz as primeiras etapas posso chegar nesse conjunto de tokens aqui.

Expande-se as contrações, caixa d'água, copo d'água, queda na água, São Miguel d'oeste, a gente faz uma expansão de contrações, mas isso já caiu um pouco em desuso no nosso português, mas no português de Portugal ele é muito utilizado. Em inglês você passa a descontração do verbo em expansão.

Remove-se então as stopwords, que são palavras sem significados, artigos, pronomes, advérbios, conjunções, etc.

Corrige-se a grafia, para evitar problemas gramaticais. Obviamente vai ter que ter um dicionário aqui na língua que está trabalhando.

E faz-se o processo de lematização e radicalização. A lematização busca a forma canônica dos termos, por exemplo o verbo estudar. A lematização ela pode ser substantivada ou verbalizada.

A radicalização simplesmente é ver o que tá igual aqui, pegar a interseção e definir. Veja que a radicalização nem sempre vai resultado em um dicionário de dado, a lematização sempre vai resultar em um dicionário de dado. A radicalização é mais adequada para busca sintática, enquanto a lematização e mais adequada para busca semântica.

Finalmente então o pós-processamento tem as etapas, geralmente de validação cruzada, tem uma série de metas, você tem que avaliar o resultado do classificador, do clusterizador, do sumarizador, ver se aquilo está adequado, e então realimentar o sistema, redefinir as métricas e os parâmetros, reavaliar o corpus textual, eu corpus pode não estar adequado, pode não ter sido limpo o suficiente, talvez a lematização tivesse sido melhor que a radicalização, a gente precisa fazer esses ajustes.

Aqui um exemplo rápido da vetorização, eu peguei três documentos muito pequenos, que são conjunto de frases, e fui aplicando aqui, vejam, "ambiente agradável e tranquilo", "comida rústica com qualidade", e "adoramos o filé à parmegiana". No documento dois, "o filé parmegiana da cidade", "ambiente agradável" e "qualidade no atendimento", e documento três "o filé a parmegiana com fritas é uma delícia".

Então aplicando a análise léxica dos corpus a gente chega nesse tipo de tokens, depois de todas aqueles processos de limpeza. Depois a gente remove as stopwords, veja que já reduziu aqui minha quantidade de tokens. Finalmente a gente faz a radicalização e chega-se a esses termos aqui. No dicionário de termos a gente fica então com 13 termos naquele conjunto de documentos.

Finalmente a gente bota isso em uma matriz e faz uma representação binária, cada coluna é um dos termos e cada linha é o número do documento, e eu faço uma representação binária para ver se o termo existe ou não em determinado documento.

A partir daí, a gente calcula o Tf e ITf para cada um dos termos e dos documentos, e finalmente o cálculo normalizado do Tf e ITf. Agora eu tenho um dataset pronto para que se rode algoritmos de Machine Learning.

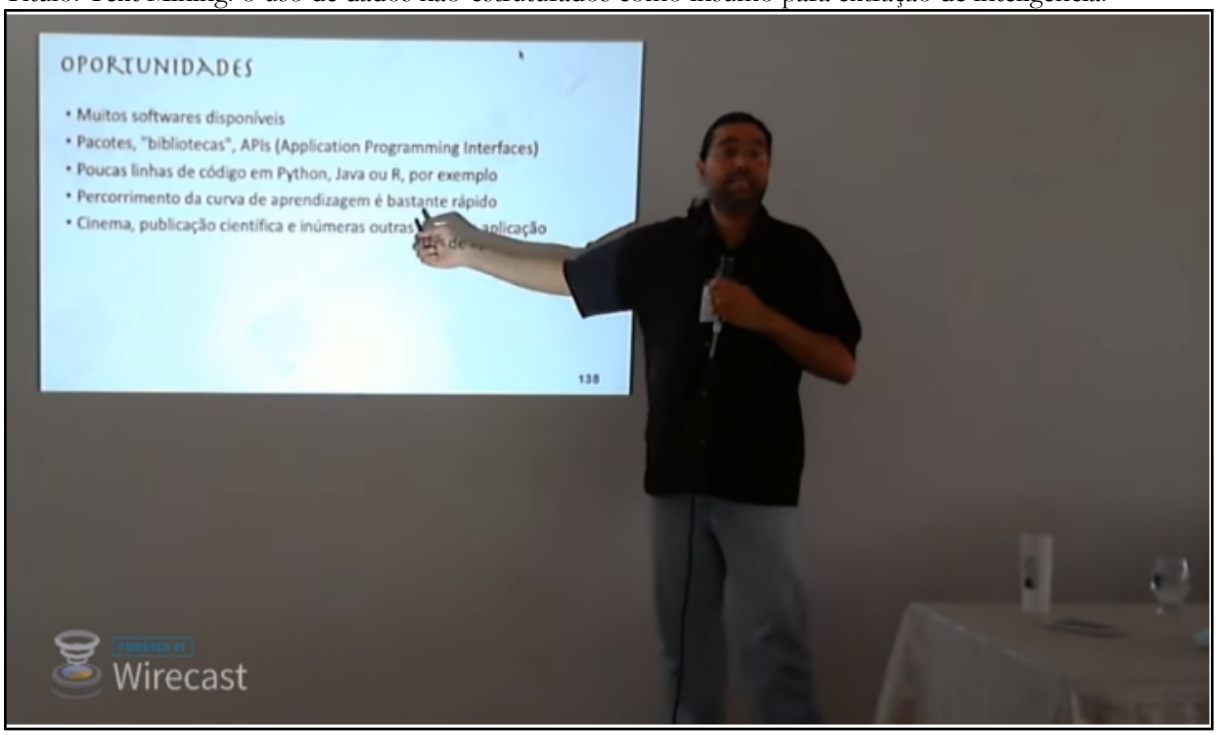
Alguns desafios da mineração de texto: minerar termos em línguas que não são inglês, é um dos grandes desafios hoje principalmente falando em termos de Brasil, a gente tem uma série de problemas que são mais complexas como relações semânticas.

E as oportunidades do Text Mining, existe uma miríade de softwares disponíveis, muitas APIs, pacotes, bibliotecas de software, códigos na linguagem Python com o uso do R a gente faz maravilhas porque está quase tudo pronto, o nosso trabalho basicamente é o trabalho de coleta, fazer web scraping, isso é uma Ad Hoc que não está pronta ainda.

A curva de aprendizagem ela é percorrida muito rapidamente, então para retomar lá do início tanto no cinema quanto na produção científica e inúmeras outras áreas elas podem ter aplicação na mineração de texto, obrigado.

Vídeo da apresentação

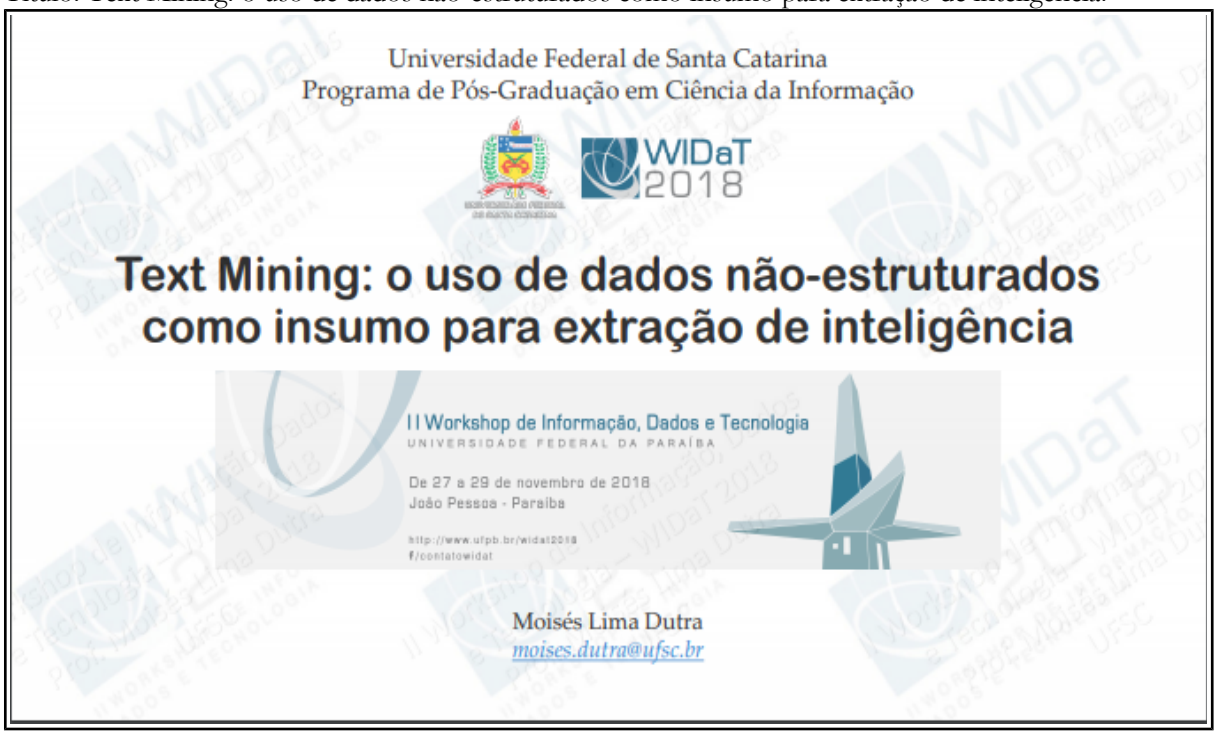
Título: Text Mining: o uso de dados não-estruturados como insumo para extração de inteligência.



Disponível em: http://dadosabertos.info/enhanced_publications/idt/video.php?id=32

Slides da apresentação

Título: Text Mining: o uso de dados não-estruturados como insumo para extração de inteligência.



Disponível em: http://dadosabertos.info/enhanced_publications/idt/presentation.php?id=32