

¿Qué es Text Mining?

Text Mining es ir de...

“La calidad de atención es muy mala. Llamé 10 veces para que me ayudaran con el armado de mi bomba para correccaminos marca Acme, y no me atendió nadie”.

...a:

Razón: Calidad de atención mala. Llamó 10 veces.

Tópico: instalación de bomba marca Acme.

Esto se puede combinar con tiempo de llamada, operador, etc para ver cuales son los productos más caros de atender.

¿Qué es Text Mining?

Text mining es extraer información interesante y no obvia de texto sin estructura, para poder encontrar patrones y tendencias de datos, asociaciones entre entidades, y reglas predictivas entre esas entidades mencionadas en el texto.

¿Para qué text mining?

A veces tenemos mucho texto y poca información, escondida...

¿Quién es autoridad en qué tema? ¿Qué temas están creciendo en importancia?

Otras tenemos mucho texto y mucha información escondida...

emails, chats, libros online, comentarios de productos ¿Qué leo? ¿Con qué se relaciona? ¿Tiene el texto un sesgo a favor o en contra de algo o alguien?

¿Por qué text mining?

- En algunos campos (p.ej. relacionados a biología) el 80% del conocimiento está en papers.
 - Humanos no escalan: Ud puede leer digamos 20 papers por semana. En ese lapso PubMed agregó 2500 abstracts.
 - De acuerdo a Gartner, hasta el 85% de información empresarial es no estructurada.
-

Buscar no es suficiente

El objetivo de Information Retrieval (IR) es ayudar a usuarios a que encuentren una respuesta a una necesidad de información, o sea maximizar precision y recall. No es tanto que la información no esté clara, sino que es muy difícil de encontrar.

El objetivo de Text Mining es el de identificar y extraer y relacionar información con mayor precisión.

Técnicas de IR se usan en text mining, p. ej. representación de documentos, clustering de documentos, análisis de citas y links.

Text Mining es difícil

Text Mining es diferente data mining tradicional porque:

- Las computadoras no pueden leer (comprender) texto.
 - El texto no tiene estructura bien definida (campos).
 - Un documento trata varios temas.
 - El significado de las palabras es ambiguo, y depende del contexto y del idioma.
 - Posible explosión combinatoria de conexiones potencialmente válidas.
-

Text Mining es difícil

- Número de atributos > 15000
 - Ruido (errores de ortografía, abreviaturas)
 - Sinónimos
 - Diferentes significados dependiendo de la función:
("claro, lo que ud quiere es..." vs. "los de colores claros son mas caros")
 - Diferentes significados dependiendo de la comunidad:
"debian soluciona tu problema" (Debian Linux) vs "debian solucionar tu problema" (Verbio "debían" sin acento).
 - Localismos
-

¿Para qué sirve?

- Detección de importancia de emails (spam)
 - Detección de comentarios ofensivos (KeepCon)
 - Clustering de documentos (www.polymeta.com, CiteseerX)
 - Creación de mapas de tópicos (info.leximancer.com)
 - Minería de tendencias, y Opiniones (forsight, SocialMetrix)
 - Extracción de información ontologías y de entidades (whalewisdom.com procesa los reportes de hedge funds)
 - Resumen de documentos (Yahoo News Digest)
 - Descubrimientos de relaciones por citas (ACM DL, Google Scholar)
-

Clustering: iboogie

The screenshot shows the iBoogie search engine interface. The browser address bar displays the URL `iboogie.com/searchtree.asp?name_query=text+mining&nar`. The search bar contains the text "text mining" and a "Search" button. To the right of the search bar is a link for "Advanced Search".

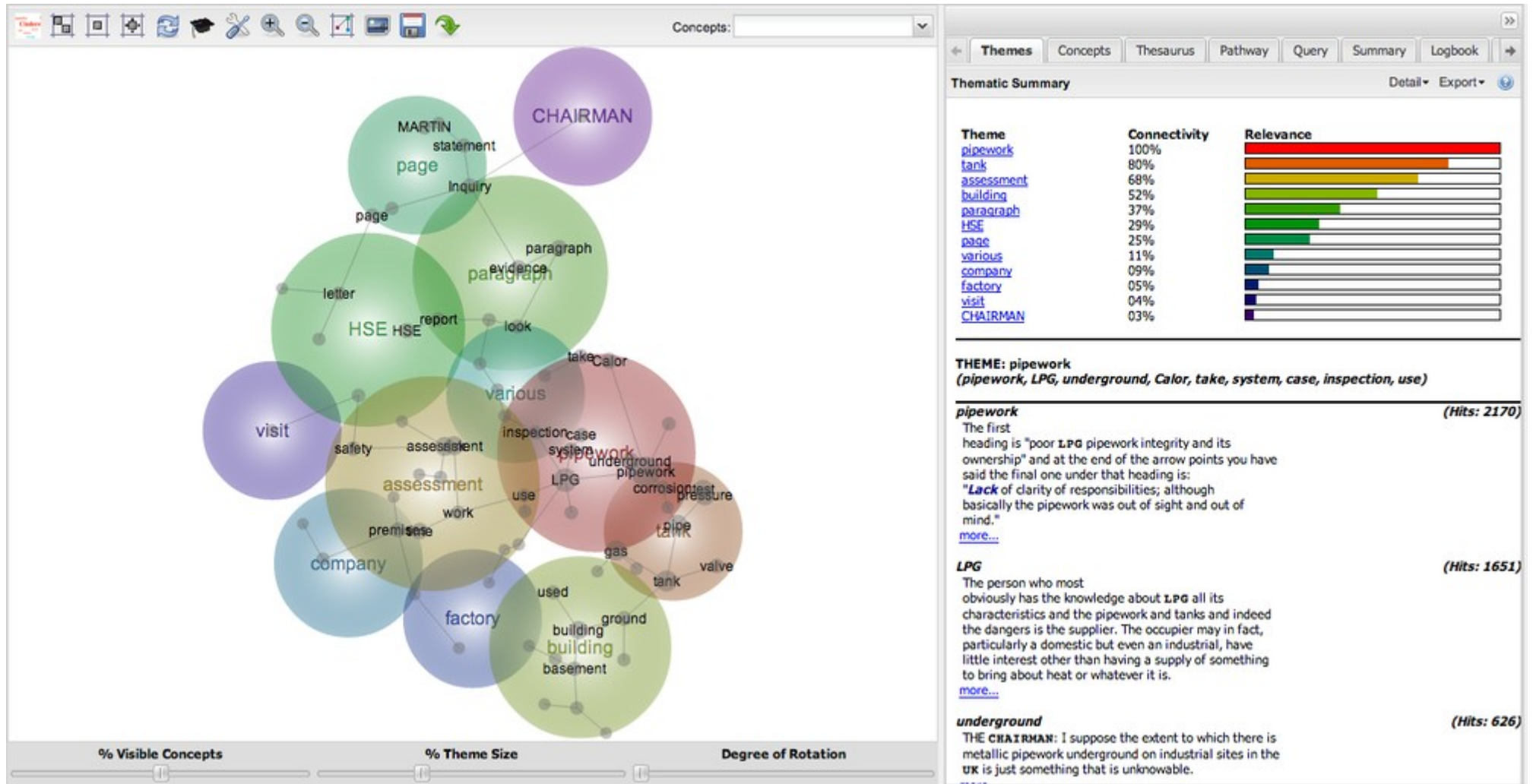
On the left side, under the "iBoogie™" logo, there is a list of categories with expandable icons (+):

- Text and data mining
- Text analytics
- Applications
- Text mining tools
- Text analysis
- Text mining software
- Mining The process
- Text mining techniques
- Documents
- Business
- Computer
 - Research
 - Technology
- University
- Library
- Text-mining
- Online text mining
- Introduction to text mining

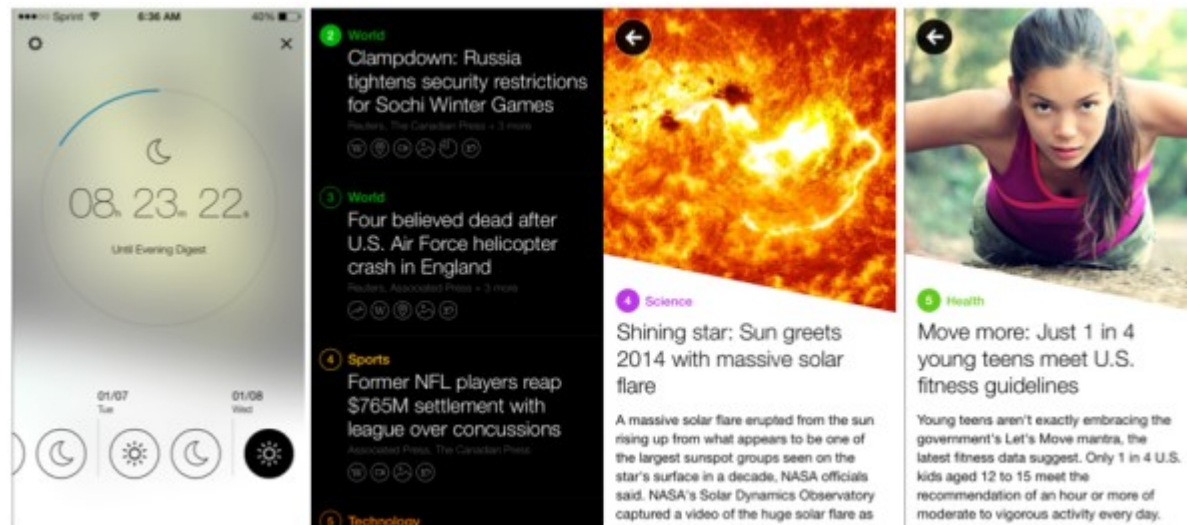
The main content area shows "100 results out of 0 found". Below this, several search results are listed:

- Text mining - Wikipedia, the free encyclopedia** ☐
Text mining, also referred to as **text data mining**, roughly equivalent to **text** analytics, refers to the process of deriving high-quality information from **text**.
http://https://en.wikipedia.org/wiki/Text_mining - MSN
- What is text mining (text analytics)? - Definition from ...** ☐
Text mining is the analysis of data contained in natural language **text**. The application of **text mining** techniques to solve business problems is called **text** analytics.
<http://searchbusinessanalytic...get.com/definition/text-mining> - MSN
- Text Mining - Statistics Textbook** ☐
Text Mining Introductory Overview. The purpose of **Text Mining** is to process unstructured (textual) information, extract meaningful numeric indices from the **text**, and ...
<http://documents.software.del...tatistics/Textbook/Text-Mining> - MSN
- Marti Hearst: What Is Text Mining?** ☐
I wrote this essay for people who are curious about the topic of **text mining** after having read the New York Times article by Lisa Guernsey (10/16/2003) or heard my ...
<http://people.ischool.berkeley.edu/~hearst/text-mining.html> - MSN
- Text Mining Software, SAS Text Miner | SAS** ☐

Mapas de temas: Leximancer



Resúmenes: Yahoo! News Digest app



Extracción de información: whalewisdom

The screenshot displays the WhaleWisdom Filer Search interface. The top navigation bar includes a search bar, 'Search Active' status, filters for '13F' and 'Other', and a user profile 'brent@whalewisd...'. The left sidebar contains navigation links for Dashboard, My Tracked Backtests, My Filer Groups (8), My Email Alerts (2), Reporting/Analytics (Consensus Reports, Combined Holdings Reports, Run Backtest, 13F Charting), and Search Tools (Filer Search, 13F Fund Performance, Investment Advisor Search, 13F Stock Screener, Form D Search). The main content area is titled 'Filer Search' and features a table of search results. The table columns are: Type of Filer, Total MV of 13F Holdings, # of 13F Holdings, Last 13F Filed, Top 10 Holdings % of Total, and WhaleScore 1-yr Avg. A 'Beat S&P 500' badge is visible. The table lists three companies: BB BIOTECH AG (WhaleScore 94), PAR CAPITAL MANAGEMENT INC (WhaleScore 90), and BAKER BROS. ADVISORS LP (WhaleScore 90). Each row shows overall performance, total value, number of holdings, top 10 holdings percentage, and last filing date. An 'EXPORT TO EXCEL' button is located in the top right corner.

| Type of Filer | Total MV of 13F Holdings | # of 13F Holdings | Last 13F Filed | Top 10 Holdings % of Total | WhaleScore 1-yr Avg |
|----------------------------|--------------------------|-------------------|----------------|----------------------------|---------------------|
| BB BIOTECH AG | \$3.09 b | 28 | 2014-12-31 | 71.33% | 94 |
| PAR CAPITAL MANAGEMENT INC | \$4.33 b | 70 | 2014-12-31 | 58.11% | 90 |
| BAKER BROS. ADVISORS LP | \$9.80 b | 117 | 2014-12-31 | 80.40% | 90 |

Extracción de información: LUIS

New feature: Composite entities are live! Start creating/editing your application to use them.

Help

The pre-built personal assistant V2 can interpret commands like "s functionality, see the [documentation page](#) for this pre-built applic

Query

levantarme a las 6 de la mañana para ir al trabajo

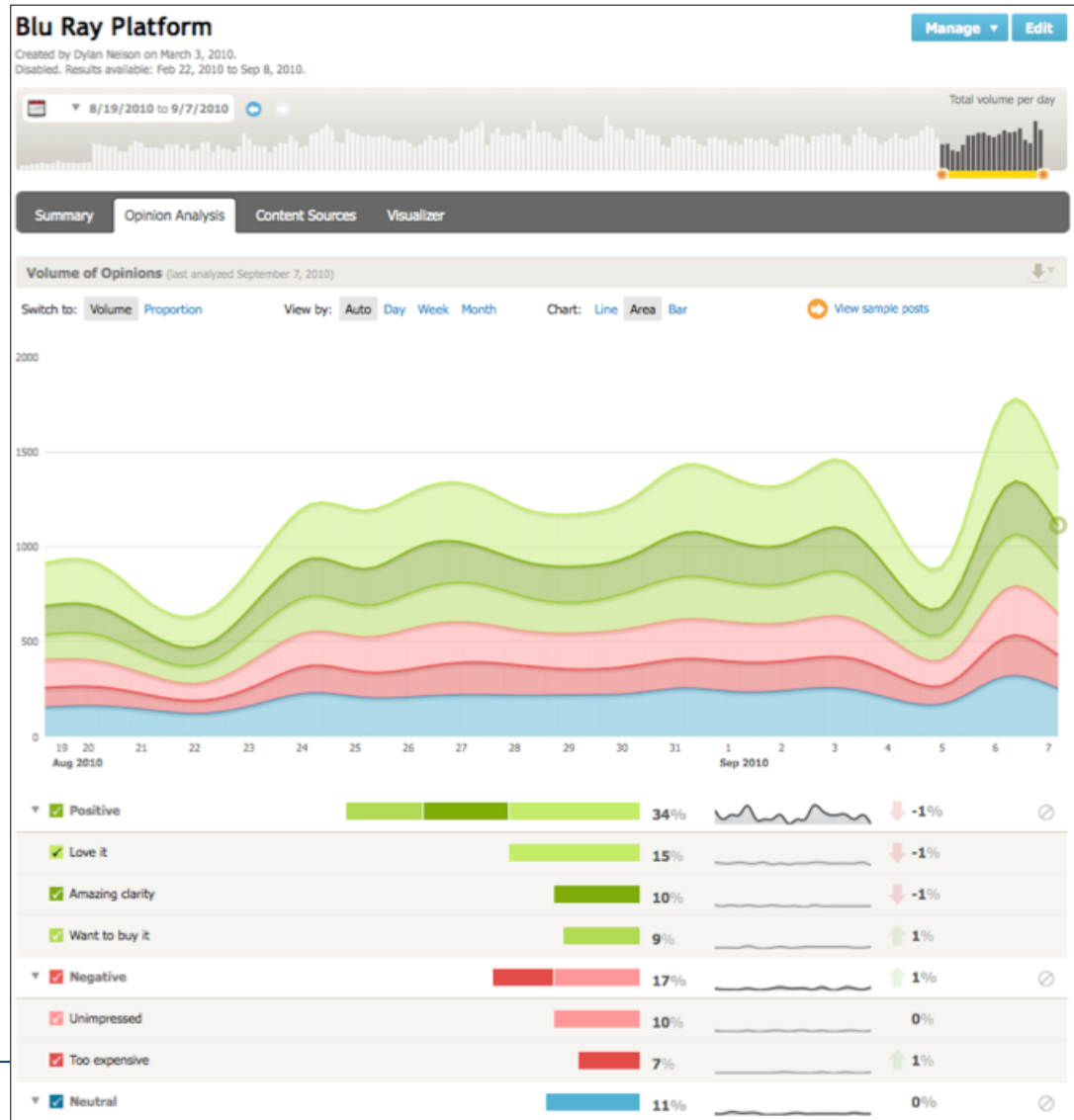
URL

<https://api.projectoxford.ai/luis/v1/application?id=cb2675e5-fbea-4f8b-89353c641a4d600&q=levantarme%20a%20las%206%20de%20la%20ma%C3>


Advanced

```
{
  "query": "levantarme a las 6 de la mañana para ir al trabajo",
  "intents": [
    { "intent": "builtin.intent.alarm.set_alarm" }
  ],
  "entities": [
    {
      "entity": "6 de la mañana",
      "type": "builtin.alarm.start_time"
    },
    {
      "entity": "ir al trabajo",
      "type": "builtin.alarm.title"
    }
  ]
}
```

Minería de opiniones: Forsight Social Media




Descubrimiento de citas: ACM DL



[SIGN IN](#) [SIGN UP](#)


[SEARCH](#)

Latent dirichlet allocation


Full Text:  [PDF](#) [SIGN IN to get this Article](#)

Authors: [David M. Blei](#) [Computer Science Division, University of California, Berkeley, CA](#)
[Andrew Y. Ng](#) [Computer Science Department, Stanford University, Stanford, CA](#)
[Michael I. Jordan](#) [Computer Science Division and Department of Statistics, University of California, Berkeley, CA](#)

Published in:
• Journal
The Journal of Machine Learning Research [archive](#)
Volume 3, 3/1/2003
Pages 993-1022
[JMLR.org](#)
[table of contents](#)





2003 Article


 [Bibliometrics](#)


- Downloads (6 Weeks): 262
- Downloads (12 Months): 3,362
- Downloads (cumulative): 16,324
- Citation Count: 2,495

Tools and Resources


 TOC Service:
[Email](#) [RSS](#) [RSS](#)

 [Save to Binder](#)

 Export Formats:
[BibTeX](#) [EndNote](#) [ACM Ref](#)

Share:
 [AddThis](#)

Tags: [algorithms](#)
[connectionism and neural nets](#)
[knowledge acquisition](#) [text analysis](#)

 [Contact Us](#) | [Switch to single page view](#) (no tabs)

[Abstract](#) [Authors](#) [References](#) [Cited By](#) [Index Terms](#) [Publication](#) [Reviews](#) [Comments](#) [Table of Contents](#)

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

Aplicaciones de Text Mining

- En organizaciones, text mining se usa para identificar expertos y relaciones entre empleados y proyectos, tecnologías y clientes (Knowledge Management).
 - En empresas, text mining se usa para hacer análisis de información de clientes (por ejemplo, de que se queja la gente en call centers)
 - En sitios web, para moderar y detectar automáticamente comentarios ofensivos.
 - En e-commerce para extraer automáticamente fichas de producto (promptcloud extrae info de páginas en Amazon.com)
 - En química y medicina, text mining se usa para identificar nuevas relaciones entre síntomas causas y tratamientos (p.ej. Entre agentes químicos y drogas).
 - En finanzas, se usa para descubrir patrones que relacionan información financiera y no financiera con el comportamiento de una empresa en el mercado.
-

Modelos básicos de representación de documentos

- Modelo Booleano
 - Modelo Vectorial
 - Modelos Probabilísticos
-

Bolsa de Palabras: Presunciones clásicas

- Cada documento se representa por una serie de términos.
 - Un término es una palabra o grupo de palabras útiles para describir el contenido del documento.
 - Todos los términos son independientes entre sí, lo que implica que puedo calcular la importancia de cada término en un documento independientemente de los otros (la independencia no es cierta, pero en la práctica funciona).
 - El peso w_{ij} de un término t_i en un documento d_j es proporcional a la importancia de ese término.
-

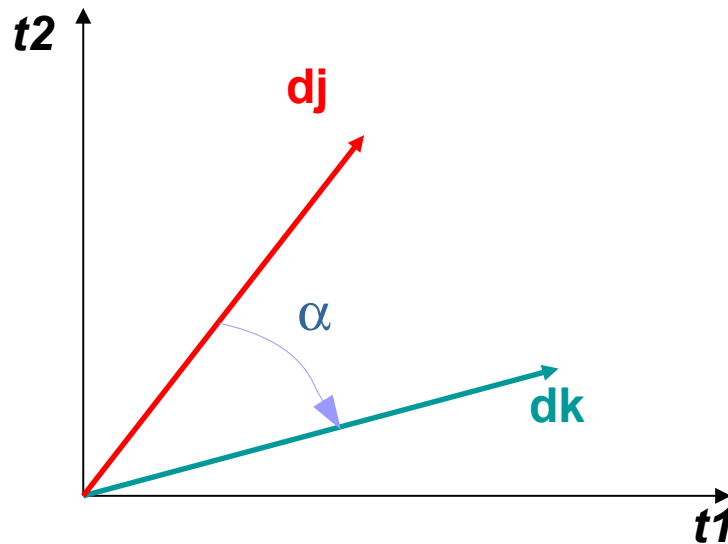
Modelo Vectorial

- Todos los términos de todos los documentos en la colección tienen un índice i único. i es siempre igual para t en todos los documentos.
 - $w_{ij} > 0$ si t_i es un término en d_j , 0 si $t_i \notin d_j$.
 - Si hay N términos en total en la colección, un documento d_j es un vector de dimensión N :
 - Documento $d_j = [w_{1j}, w_{2j}, w_{3j}, \dots, w_{Nj}]$
 - La mayoría de los w_{ij} van a ser 0 (vectores raros)
-

Modelo Vectorial

- Los D documentos son vectores en un espacio N dimensional.
 - Como t_i es independiente de t_j , entonces los vectores unitarios $t_1 \dots t_N$ son linealmente independientes (forman una base del espacio).
 - Tanto los documentos como las consultas son vectores dentro del espacio.
-

Similaridad entre Documentos



$$\begin{aligned} \text{sim}(d_i, d_k) &= \cos(\alpha) = [\text{vec}(d_i) \bullet \text{vec}(d_k)] / |d_i| \times |d_k| \\ &= [\sum (w_{ij} \times w_{ik})] / |d_i| \times |d_k| \end{aligned}$$

como $w_{ij} \geq 0$ y $w_{ik} \geq 0$, $0 \leq \text{sim}(d_i, d_k) \leq 1$

- *Dos documentos se parecen si tiene algún término en común (no necesariamente con igual importancia).*
-

Modelo Vectorial

- ¿Cómo calcula el peso de W_{ij} ?
 - El más simple: Si $t_i \in d_j$ entonces $W_{ij}=1$ (modelo booleano).
 - Sin embargo, un buen valor de W_{ij} debe tomar en cuenta:
 - Que tan importante es t_i en el documento d_j
 - Qué tan bien describe t_i a d_j en particular.
-

Importancia de un Término

- **Que tan importante es t_i en el documento d_j :** La frecuencia del término $tf(i,j)$ es proporcional a la importancia de t_i en d_j .
 - **Que tan bien t_i describe a d_j en particular:** Cuanto menos aparezca t_i en D , más específico es t_i .
 - idf = inverse document frequency
-

Midiendo la importancia de un Término

- **ti** = término nro. i en la colección de docs.
 - **N** = # de documentos en la colección.
 - **ni** = # de documentos que contienen a ti .
 - **$tf(i,j)$** = frecuencia de ti en el doc dj .
 - **tf nomalizado** = $tf(i,j)/\max(tf(p,j))$, donde p es cada término en dj .
 - **$idf(i)$** = $\log(N/ni)$ Log se usa para que los rango de valores de tf y de idf sean mas cercanos.
-

Importancia de un Término

El peso de w_{ij} mas usado es $w_{ij} = tf(i,j) \times idf(i)$

Para las consultas, una manera común de pesar los terminos es:

$$w_{iq} = (0.5 + [0.5 * freq(i,q) / max(freq(i,q))]) * log(N/ni)$$

El modelo vectorial es usualmente tan bueno como los otros modelos, y es rápido para computar, aunque sin una base teórica.

Mejorando una búsqueda

- Feedback de relevancia son técnicas usadas para modificar las consultas en base a lo que el usuario indica que es interesante.
- Un método simple es:
 - 1) El usuario ejecutan un query q , e identifica documentos relevantes (D_r) y no relevantes (D_{nr}) en la lista de resultado de búsqueda.
 - 2) El sistema crea un nuevo query qe utilizando términos de D_r y D_{nr} para modificar al query q , y retorna otra lista de resultados.
- El método más simple es el método de Rocchio: (α , β y γ son parámetros):

$$\vec{q}_e = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{d_r \in D_r} \vec{d}_r - \frac{\gamma}{|D_{nr}|} \sum_{d_{nr} \in D_{nr}} \vec{d}_{nr}$$
