# Aplicaciones de la Teoría de Redes al Procesamiento del Lenguaje Natural

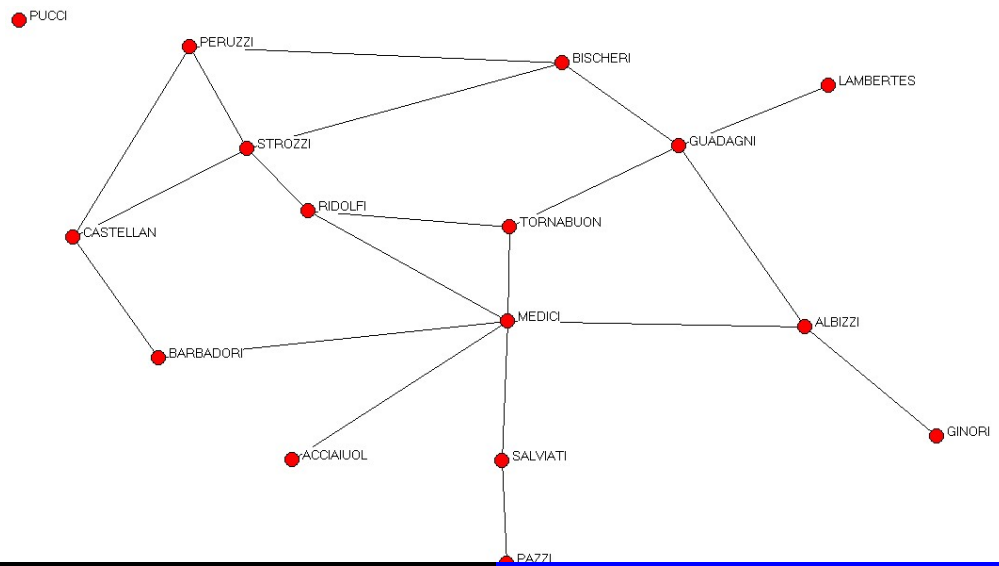
Alvaro J. Riascos Villegas

Febrero, 2016

# Contenido

# Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Padgett, J.F y C.K. Ansell (1993). Robust action and the rise of the Medici, 1400-1434.
- Red de matrimonios entre familias (cada enlace representa un matrimonio entre miembros de dos familias).



# Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Basados en la riqueza y poder político es difícil explicar como los Medici surgieron como una familia tan importante (la familia Strozzi tenía más riqueza y poder político, sin embargo fueron opacados por los Medici).
- La estructura de relaciones puede ser un determinante.
- Si comparamos con cuántas familias se encuentra una familia específica relacionada y comparamos entre ellas, los Medici sobresalen (3 a 2).
- Una relacion de cercania resulta más sugestiva.

# Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Sea $P(ij)$ el número de caminos más cortos que conectan una familia $i$ con $j$. Sea $P_k(ij)$ el número de estos caminos que incluyen a la familia $k$.
- Por ejemplo si $i$ =Barbadori, $j$ =Guadagni, entonces $P(ij) = 2$. Si $k$ =Medici entonces $P_k(ij) = 2$ mientras que si $k =$ Strozzi o Albizzi $P_k(ij)$ es cero o uno respectivamente.

# Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Si calculamos una medida de importancia (betweenness Freeman) de cada familia $k$ como:

$$\sum_{i,j:i\neq j,k\in\{i,j\}} \frac{\frac{P_k(ij)}{P(ij)}}{(n-1)(n-2)/2} \tag{1}$$

donde $\frac{P_k(ij)}{P(ij)} = 0$ si no hay caminos entre $i$ y $j$. El coeficiente $(n-1)(n-2)/2$ es el número máximo de pares de familias que incluirían a la familia $k$.

# Ejemplos de Redes: Familia Medici, Florencia 1400

- Esta medida de poder para los Medici es 0.522. Esto significa que los Medici están en más de la mitad de los caminos más cortos entre todos los caminos más cortos entre cada par de familias.
- Este mismo cálculo para los Strozzi es 0.103. El segundo más alto es los Guadagni que es 0.255.
- En este sentido los Medici estaban mejor posicionados que cualquier otra familia.
- Esta estructura es endógena? Es óptima?

# Ejemplos de Redes: Amistades y romances en estudiantes secundaria

- Datos de $90,000$ estuadiantes de la encuesta Add Health entrevistados en los años 90.
- A los estudiantes se les preguntaba con quién habian tenido relaciones romanticas en los últimos seis meses.

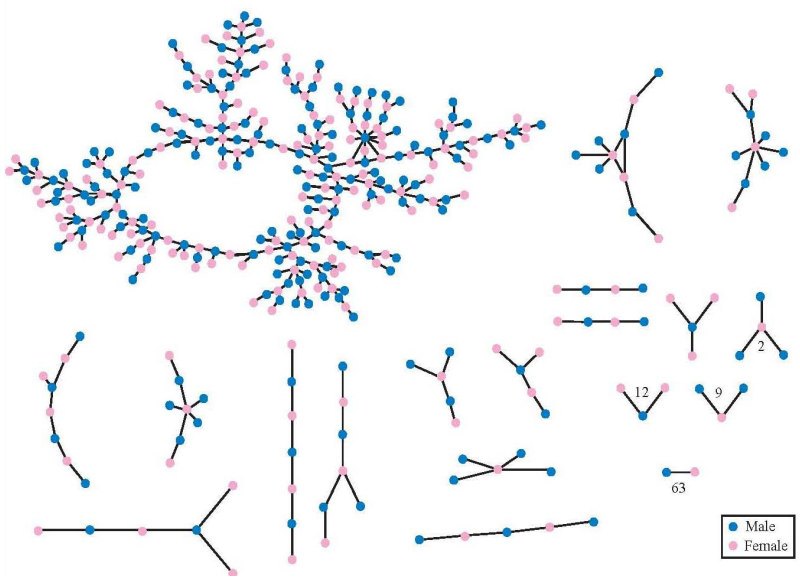



Figure 1.2: A Figure from Bearman, Moody and Stovel [47] based the Add Health Data Set. A Link Denotes a Romantic Relationship, and the Numbers by Some Components Indicate How Many Such Components Appear.

# Ejemplos de Redes: Formación aleatoria de redes

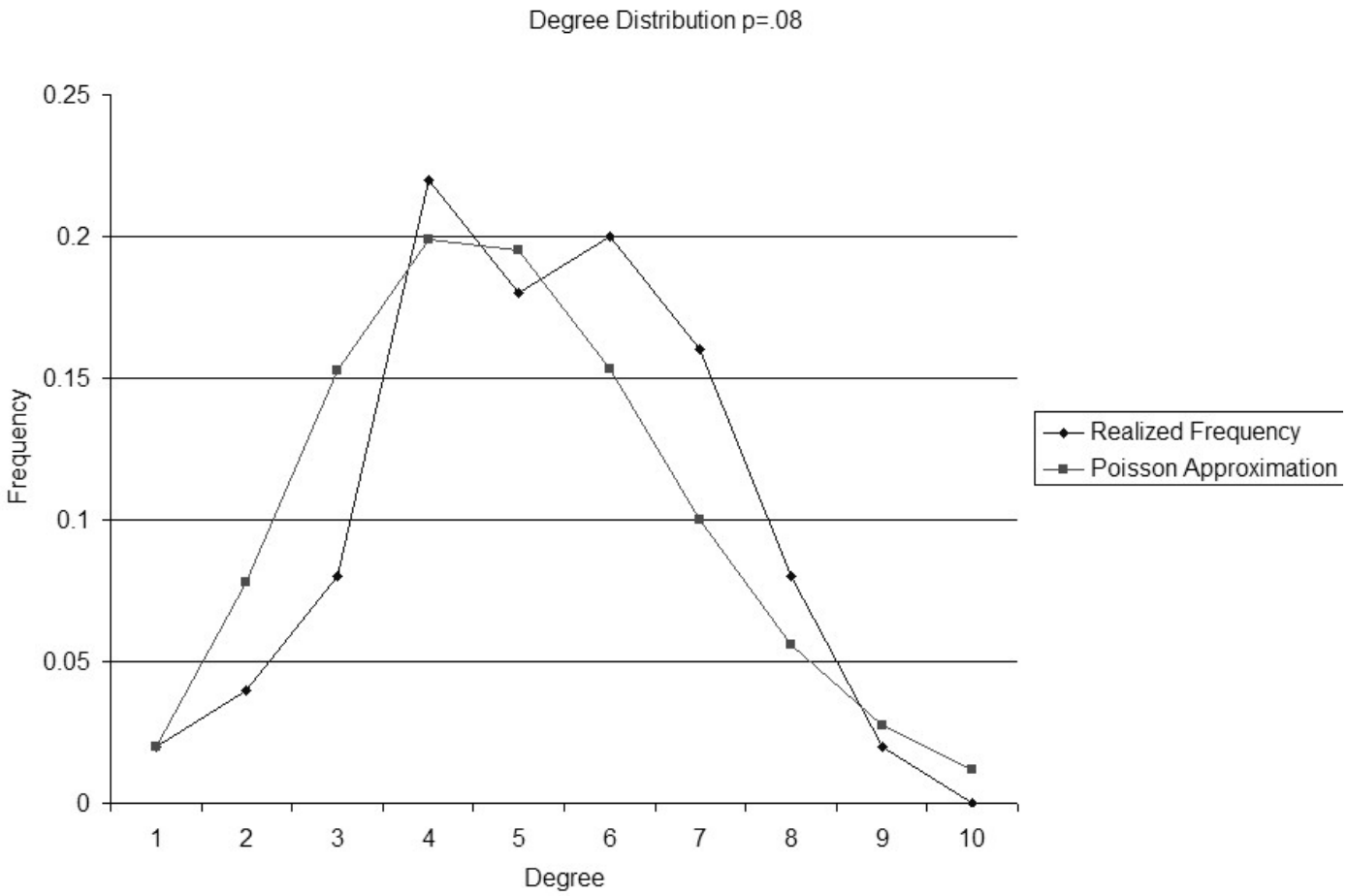



Figure 1.7: Frequency Distribution of a Randomly Generated Network and the Poisson Approximation for a Probability of .08 on each Link

# Contenido

# Redes do Coocurrencia

- Son redes que se construyen usando como nodos palabras (o solamente las que pertenecen algun tipo de palabra) y creando un enlace entre ellas cuando aparecen juntas en una ventana de N-palabras (puede ser en una frase o parágrafo).
- Pueden ser redes dirigidas o no y con pesos (usando la verosimilitud, etc).
- Ferrer-i-Cancho y Sole (2001) muestran que este tipo de redes tienen características de mundos pequeños. Diametro de aprox. 2,65 y coeficiente de clustering 0,5 y distr. libre de escala.
- Usan British National Corpus ($n = 470{,}000$, $m = 170mm$).

# Redes do Coocurrencia

- Son redes que se construyen usando como nodos palabras (o solamente las que pertenecen algun tipo de palabra) y creando un enlace entre ellas cuando aparecen juntas en una ventana de N-palabras (puede ser en una frase o parágrafo).
- Pueden ser redes dirigidas o no y con pesos (usando la verosimilitud, etc).
- Ferrer-i-Cancho y Sole (2001) muestran que este tipo de redes tienen características de mundos pequeños. Diametro de aprox. 2,65 y coeficiente de clustering 0,5 y distr. libre de escala.
- Usan British National Corpus ($n = 470{,}000$, $m = 170mm$).

# Redes do Coocurrencia

- Son redes que se construyen usando como nodos palabras (o solamente las que pertenecen algun tipo de palabra) y creando un enlace entre ellas cuando aparecen juntas en una ventana de N-palabras (puede ser en una frase o parágrafo).
- Pueden ser redes dirigidas o no y con pesos (usando la verosimilitud, etc).
- Ferrer-i-Cancho y Sole (2001) muestran que este tipo de redes tienen características de mundos pequeños. Diametro de aprox. 2,65 y coeficiente de clustering 0,5 y distr. libre de escala.
- Usan British National Corpus ($n = 470{,}000$, $m = 170mm$).

# Redes do Coocurrencia

- La distribución del grado sigue una power law con parámetro $-1{,}5$, $-2{,}7$. Esta regularidad empírica sirve para validar una teoría: existe un kernel lexicon del inglés que comparten la mayoría de personas (50m palabras) y una lista de palanbra menos usadas.

# Redes do Coocurrencia: Ejemplos

- Ferrer-i-Cancho y Sole (2001) muestran que este tipo de redes tienen características de mundos pequeños. Diametrode aprox. 2,65 y coeficiente de clustering 0,5 y distr. libre de escala.
- Usan British National Corpus ($n = 470{,}000$, $m = 170mm$).
- La distribución del grado sigue una power law con parámetro $-1{,}5$, $-2{,}7$. Esta regularidad empírica sirve para validar una teoría: existe un kernel lexicon del inglés que comparten la mayoría de personas (50m palabras) y una lista de palanbra menos usadas.

# Redes do Coocurrencia: Ejemplos

- Masucci y Rodgers (2006). Palabras adyacentes en la novela de Orwell 1984.
- Encuentran evidencia fuerte de power laws.

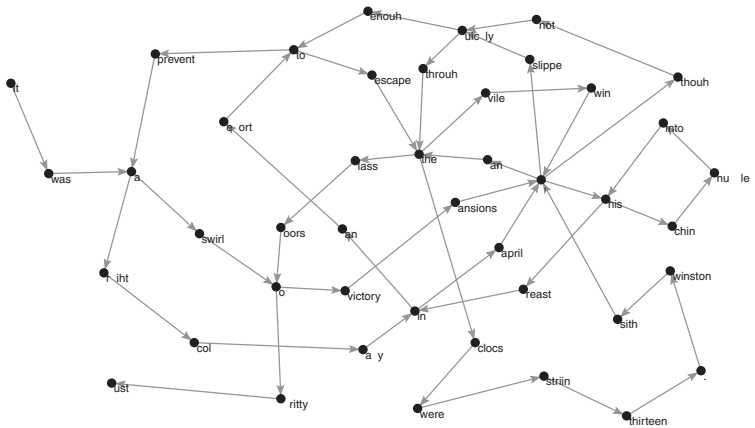



Figure 4.2. The graph corresponding to the beginning of *1984*.

# Contenido

# Ejemplos de Redes: Semánticas

- Steyvers y Tenenbaum (2005): Experimento con seis mil personas. Se les dan 500 palabras y ellos declaran si hay alguna asosicación entyre pares.

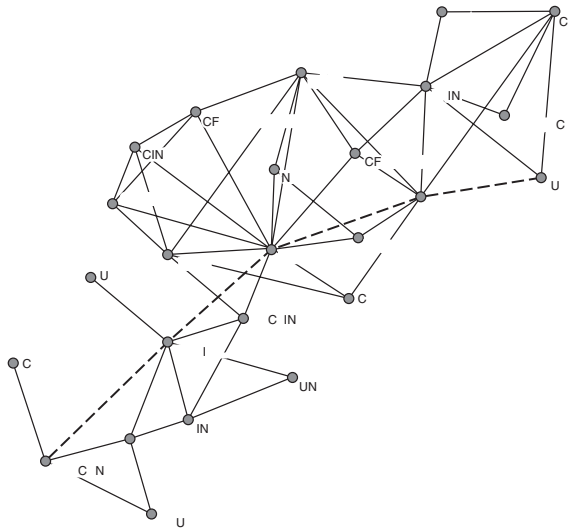



Figure 4.5. Free-word-association network (Steyvers and Tenenbaum 2005b).

# Ejemplos de Redes: Semánticas

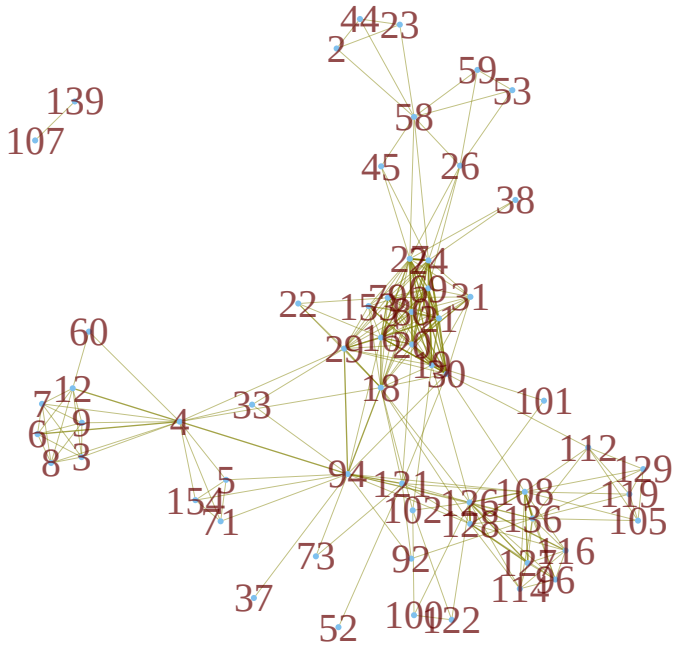
Table 4.2. *Properties of semantic networks.*

| Property | Associative | Roget | WordNet |
|---|---|---|---|
| $C$ | 0.186 | 0.875 | 0.0265 |
| $C_r$ | $4.35 \times 10^{-3}$ | 0.613 | $1 \times 10^{-4}$ |

The comparison is between the actual clustering coefficient and the clustering coefficient of a random graph of the same size.

# Contenido

# Clasificación de textos: Grafo de Documentos

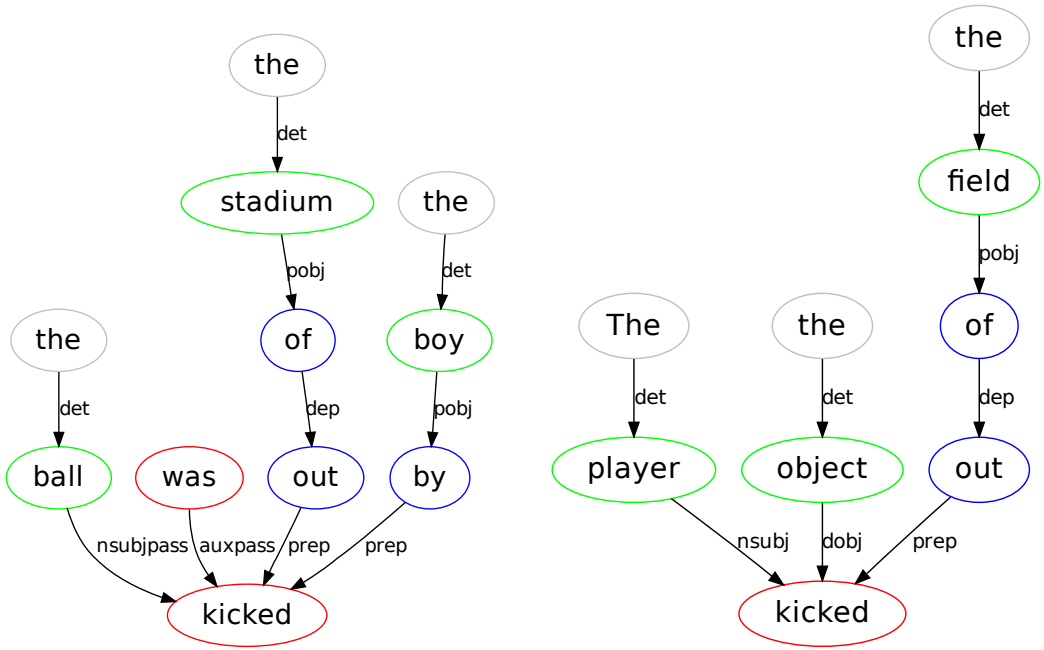


# Clasificación de textos

- An undirected bipartite graph is a triple G = (D, W, E) where D, W are two sets of vertices and E is the set of edges. Take D as the set of documents, W as the set of words they contain and an edge exists if word w occurs in document d. Note that the edges are undirected.
- In this model, there are no edges between words or between documents.
- Dhillon (2001) utiliza un algoritmo de particion de redes bipartitas.
- Tiene la ventaja de que usa información sobre las palabras y documentos para hacer la clasificación

# Copia

- Detección de plagio.
- The concrete task is to implement a graph edit distance algorithm, which calculates similarity between two graphs.
- The algorithm is based on calculating the number of edit operations needed to transform one graph into another (Riesen and Bunke, 2009).




*The ball was kicked out of the stadium by the boy*

*The player kicked the object out of the field*

Figure 1.1: Example of dependency graphs

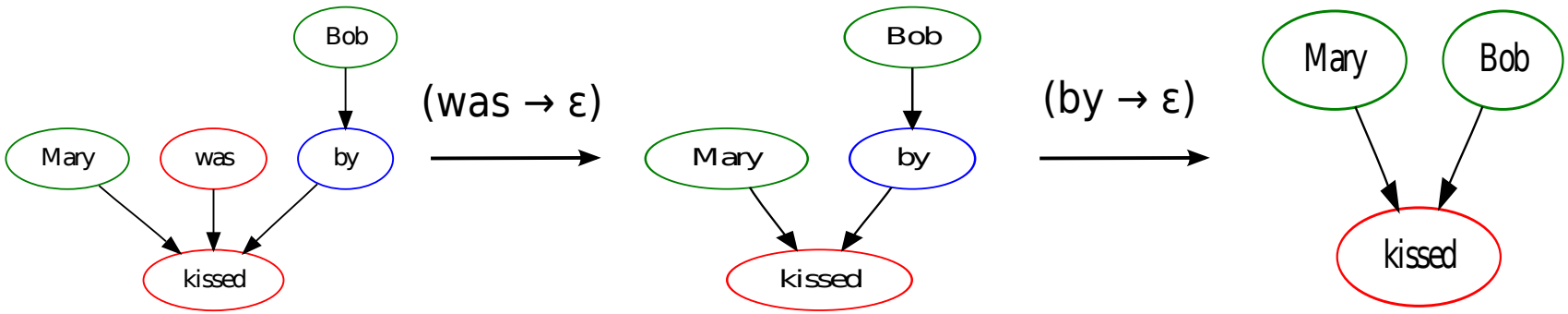



Figure 2.1: An example of edit operations for two graphs

# Contenido

# Construcción de Redes

- Red de términos.
- Red de documentos.
- Red bipartita de términos y documentos.

# Contenido

# Resumenes de Textos

- Grafos estocásticos.
- RT se basa en la relevancia de ciertes frases.
- LexRank se basa en la centralidad las frases en un grafo de frases (i.e., PageRank).
- La técnica también puede ser utilizada para tareas como: clasificación de entidades nombradas, adjunción de proposciones a frases y clasificación de textos.

# Resumenes de Textos

- La sintetización de textos es típicamente orientada por el tópico de interes.
- El problema que aquí se resuelve es de sintesis de múltiples documentos de tópicos genérico (sin especificar).
- Existen dos grande formas de hacer resumenes: extraer subconjuntos de frases de los documentos o abstracta, en la que se parafrasea las frases (típico de los resumenes humanos).
- Solamente la primera forma ha alcanzado resultados satisfactorios.

# Medidas de la Importancia de Frases: Basadas en centroides

- En un grafo de múltiples documentos se idendifican clusters.
- La centralidad de una frase se mide con base en la centralidad de las palabras que contiene.
- Para adeterminar la centralidad de las palabras se construye un pseudodocumento a partir de cada cluster (se llama centroide).
- El centroide del cluster se construye de palabras que tiene un valor mínimo de $tf \times idf$ donde $tf$ es la frecuencia de los términos en el cluster de documentos y $idf$ es la inversa de la frecuencia del término en un corpus de documentos más grande pero afín a los documentos en cuestión.
- Las frases que tengas más palabras del centroide se considera son cosideradas centrales (del cluster?).
- Esta técnica ha sido exitosa y es la base de sistemas de sintetización de multidocumentos: www.newsinessence.com

# Medidas de la Importancia de Frases: Basadas en medidas de centralidad

- Todas la propuestas a continuación se basan en el concepto de prestigio en redes.
- La hipótesis es que las frases que son similares a muchas de las frases en el cluster son mas relevantes.
- Para esto tenemos que definir similaridad entre frases y segundo la centralidad de una frase dada la similaridad a otras frases.

# Similaridad de frases

- Cada frase se representa como un vector $N - dimensional$ donde $N$ es el número de palabras posibles del lenguaje en consideración.
- Cada palabra en una frase se le asocia un número en el vector $N - dimensional$: Frecuencia del término en la frase $\times$ idf.
- La similitud entre dos frases se define como (el coseno entre dos vectores: idf-modified-cosine):

$$s(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sum_{xi \in x} (tf_{x_i,x} idf_{x_i})^2 \sum_{yi \in y} (tf_{y_i,y} idf_{y_i})^2} \qquad (2)$$

# Similaridad de frases

- Un cluster de documentos se puede representar por una matriz de de similitud (coseno) donde cada entrada es la similitud entre la pareja de frases.
- Esta matriz se puede utilizar para representar un grafo con pesos.
- Se puede escoger un umbral para disminuir el número de elementos positivos de la matriz y también olvidarnos de que es un grafo por pesos.

# Centralidad de frases

- Grado
- Prestigio (eigenvector centrality y LexRank)
- Considere la siguiente definición de centralidad de un nodo $u$:

$$p(u) = \sum_{v} \in adj[u] \frac{p(v)}{deg(v)} \tag{3}$$

  Sea $B$ la matriz que se obtiene de la matriz de adjacencia del grafo de similitud de frases normalizando cada entrada por la suma de la fila.
- $B$ representa una matriz de transición de una cadena de Markov.

# Centralidad de frases

- $B$ representa una matriz de transición de una cadena de Markov y obsérvese que $p^T B = p^T$. Es decir $p$ corresponde a una distribución estacionaria de $B$: $lim_{n \to \infty} 1^T r$.
- No importa la distribución del estado inical, asintóticamente la distribución del estado es la distribución estacionaria.
- Para garantizar la existencia de esta distribución estacionaria es necesario que $B$ sea irreducible (cada estado puede ser alcanzado con probabilida positiva despues de suficientes transiciones) y aperiodico (siempre existe la posibilidad de salir cambiar de estado).
- En estas condiciones el teorema de Perron - Frobenius garantiza la existencia de uan única distribución estacionaria.

# Centralidad de frases

- Para lograr esto se hace la siguinete modificación de la centralidad de cada nodo (i.e., PageRank):

$$p(u) = \frac{d}{N} + (1 - d) \sum_{v} \in adj[u] \frac{p(v)}{deg(v)} \tag{4}$$

  donde $N$ es el total de nodos en un grafo y $d$ es un factor de dumping (típicamente entre 0.1 y 0.2)
- La aplicación de esta forma de calcular la centralidad de cada nodo al grafo de similitud entre frases se denomina LexRank.
- Una versión alterntaiva se basa en la versión del grafo de similitud por pesos.

# Contenido

# Aplicación

- DUC 2003 y 2004 contiene 30 y 50 clusters de documentos de noticias (todos en inglés).
- Para evaluación se uso la métrica ROUGE basada en $n - gram$ coocurrencia: reporta puntajes para 1,2,3,4 - gramas de coincidencia entre los resumenes del modelo y el resumen.
- Implementaron la metodología en un software disponibl en www.summarization.xom
- Se sabe que el puntaje de los 1-gramas tiende a coincidir con las evaluaciones de humanos.