3º.Ciclo de Encontros : 4 x PLN

Redes Complexas no Processamento de Língua Natural do Português

Osvaldo N. Oliveira Jr.,
NILC, Universidade de São Paulo,
28/05/2026

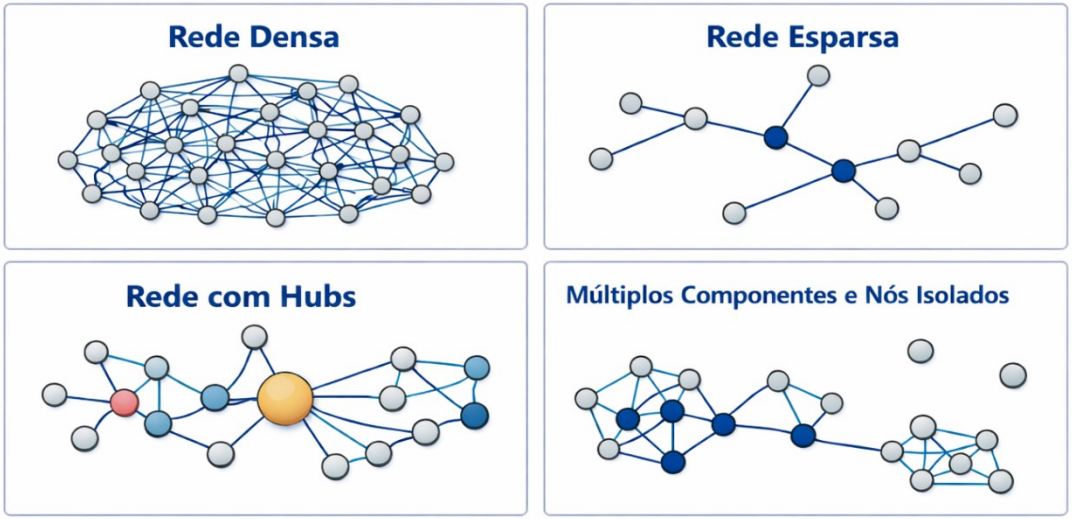
Redes Complexas no Processamento de Língua Natural do Português

Vitor H. B. D. Santi; Lucas D. V. Figueiredo; Diego R. Amancio; Maria Cristina F. Oliveira; Osvaldo N. Oliveira Jr.

IFSC-USP • ICMC-USP • NILC

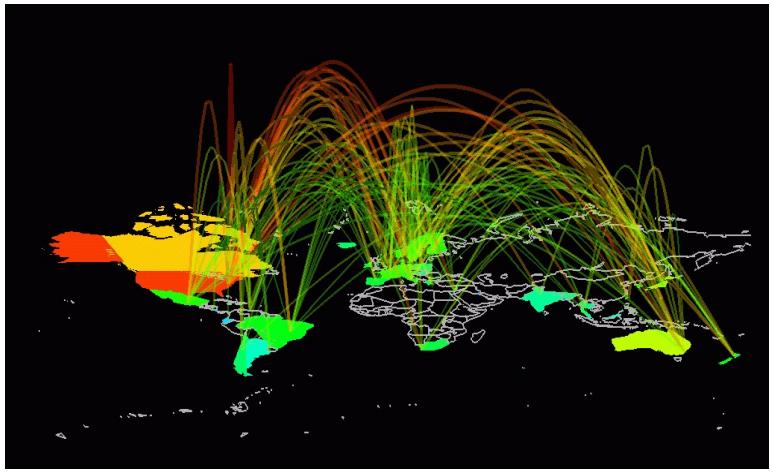
Redes Complexas

Wikipedia



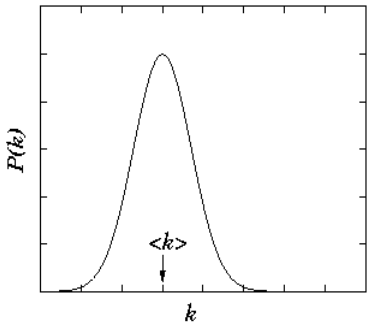
Redes complexas apresentam padrões topológicos diversos

Internet Backbone

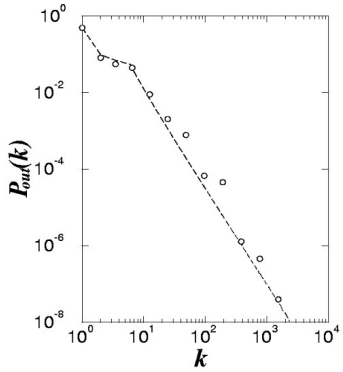


Barabási, Sci. American, 2003

Scale-free network



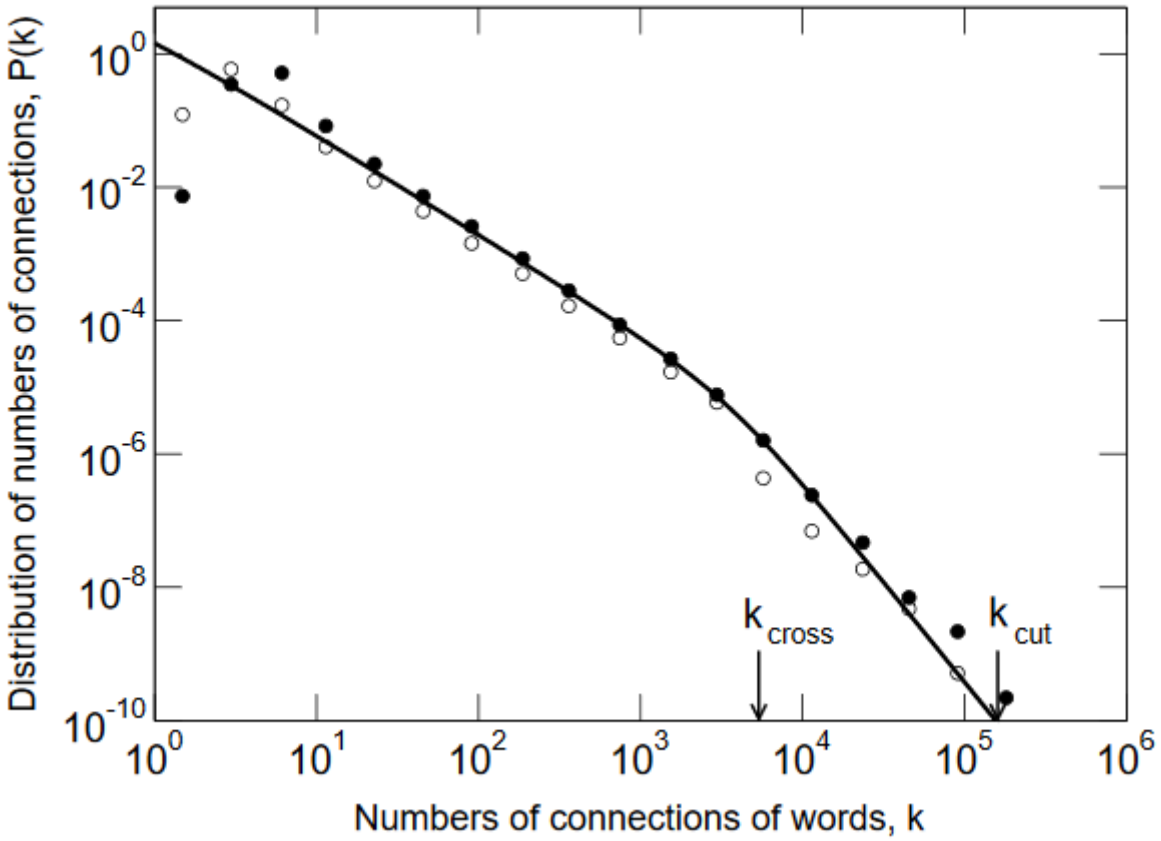
Expected



Found

Applications of Complex Networks, Costa et al., Adv. Phys., 2011

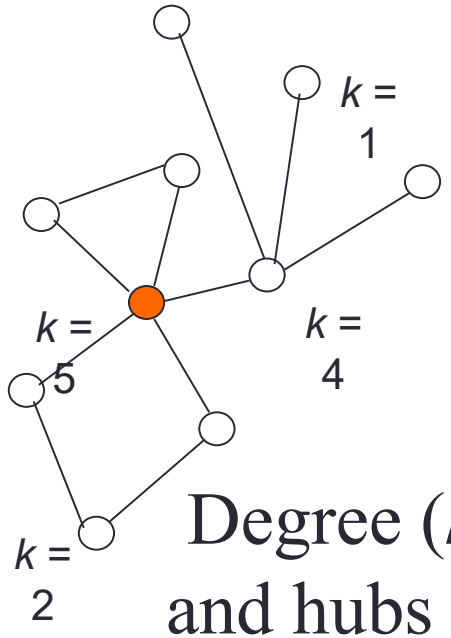
Text as complex network



Web of words is a scale-free network

Dorogovtsev and Mendes, *Advances in Physics*, 2002

Treating systems as networks

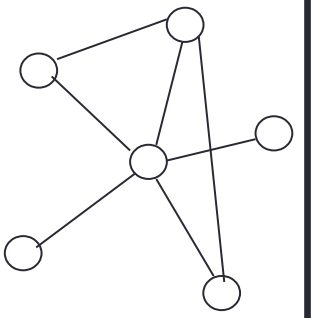


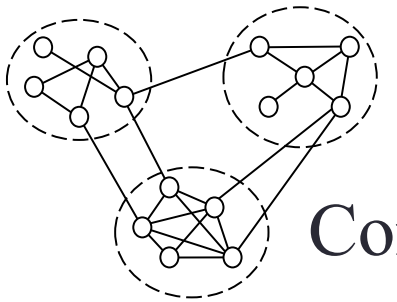
Degree (k) and hubs

Cluster coefficient

$$C = \frac{e_c}{e_T} = \frac{2e_c}{(n)(n-1)}$$

$$= \frac{(2)(2)}{(5)(5-1)} = 0.2$$





Communities

Other metrics:

Distances, shortest paths, communities, borders, accessibility, hierarchical degrees, centrality, node activity ..

More than 100 have been used for topology and dynamics

Summarization

Pardo et al., Modeling and evaluating summaries using complex networks, Lecture Notes in Artificial Intelligence, 2006.

Antiqueira et al; A complex network approach to text summarization; Information Sciences, 2009.

Amancio et al.; Extractive summarization using complex networks and syntactic dependency, Physica A, 2012.

Evaluation of Machine Translation

Amancio et al.; Complex networks analysis of manual and machine translations, International J. Mod. Phys. C, 2008.

Amancio et al.; Using metrics from complex networks to evaluate machine translation, Physica A, 2011.

The Voynich

Amancio et al;; Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript, PLOS One, 2013.

IFSC/NILC

Language analysis – including text quality evaluation

Antiqueira et al.; Strong correlations between text quality and complex networks features, Physica A, 2007.

Amancio et al.; Using complex networks to quantify consistency in the use of words, J. Stat. Mech., 2012.

Amancio et al.; Complex networks analysis of language complexity, Eur. Phys. Lett., 2012.

Amancio et al.; Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts, Physica A, 2012.

Amancio et al; Unveiling the relationship between complex networks metrics and word senses, Europhys. Lett., 2012.

Language as sensors

Dos Santos et al.; Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts, ACL, 2017

Scientometrics

Amancio et al.; Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index, J. Informetrics, 2012.

Amancio et al.; On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks, Eur. Phys. Lett., 2012.

Amancio et al.; Using complex networks concepts to assess approaches for citations in scientific papers, Scientometrics, 2012.

Silva et al.; Quantifying the interdisciplinarity of scientific journals and fields, J. Informetrics, 2013.

Amancio et al.; Topological-collaborative approach for disambiguating authors' names in collaborative networks, Scientometrics, 2015.

Authorship Identification

Amancio et al., Comparing intermittency and network measurements of words and their dependence on authorship, New J. Phys., 2011.

Amancio et al.; Identification of literary movements using complex networks to represent texts, New J. Phys., 2012.

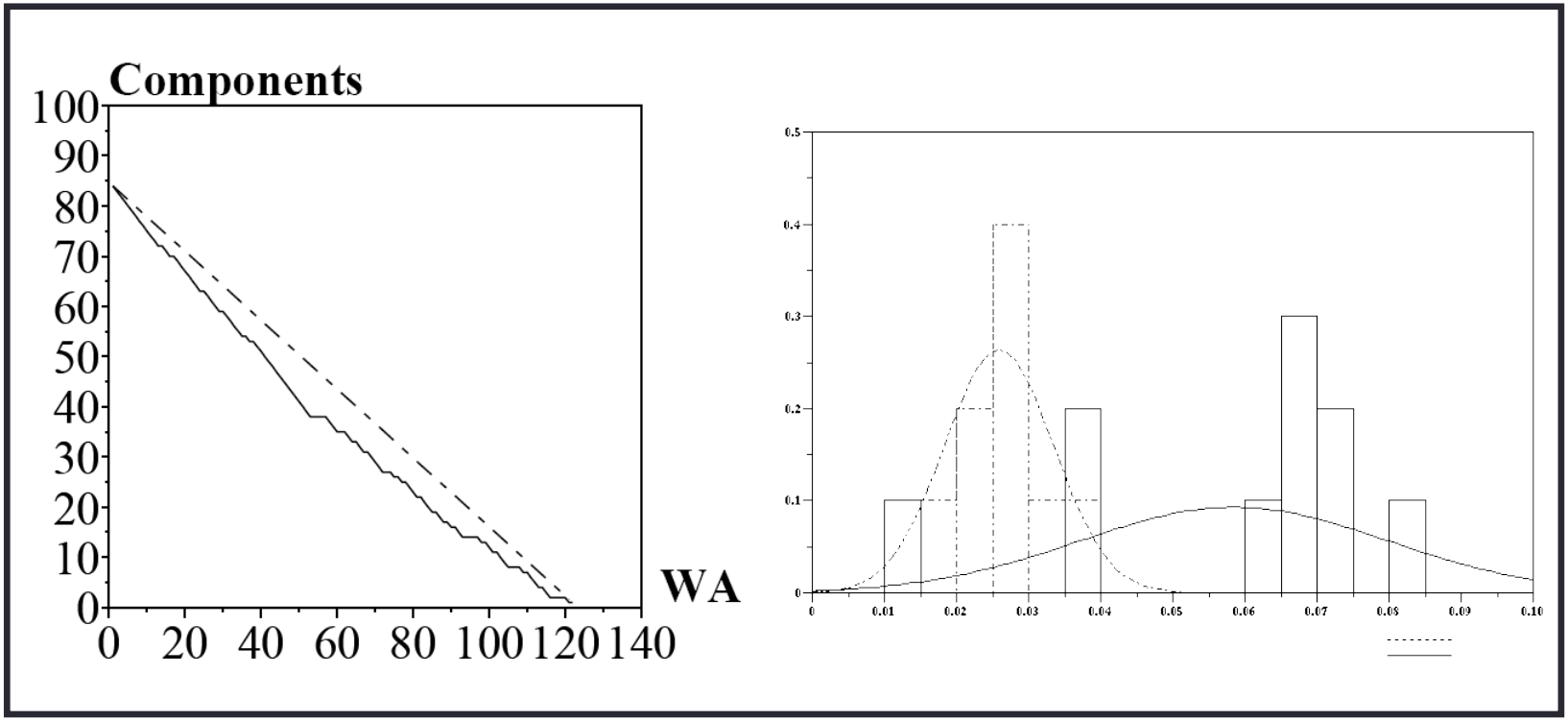
Akimushkin et al.; Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks, PLOS One, 2017.

Akimushkin et al.; On the role of words in the network structure of texts: application to authorship attribution, Physica A, 2018.

Semi-automated Surveys

Silva et al.; Using network science and text analytics to produce surveys in a scientific topic, J. Informetrics, 2016.

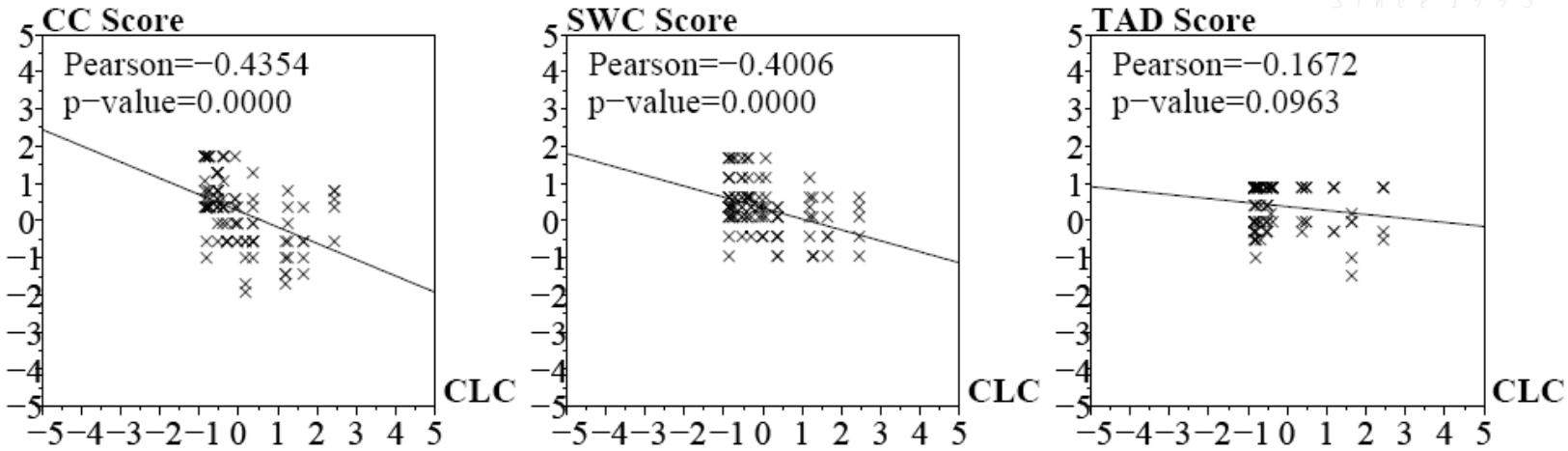
Dynamics of a network



Dynamics may be indicative of text quality

Antiqueira et al., *Physica A*, 2007

Network features



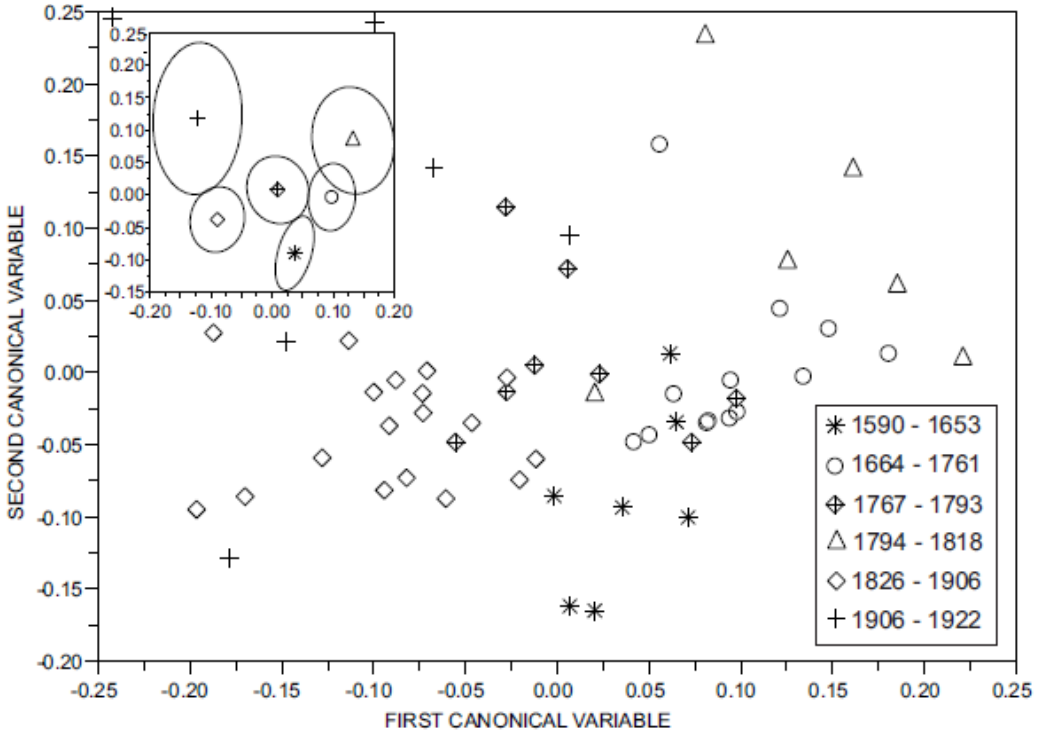
Human evaluation with regard to coherence and cohesion (CC), standard writing convention (SWC) and topic adherence (TAD) correlates well with the clustering coefficient

Antiqueira et al., *Physica A*, 2007

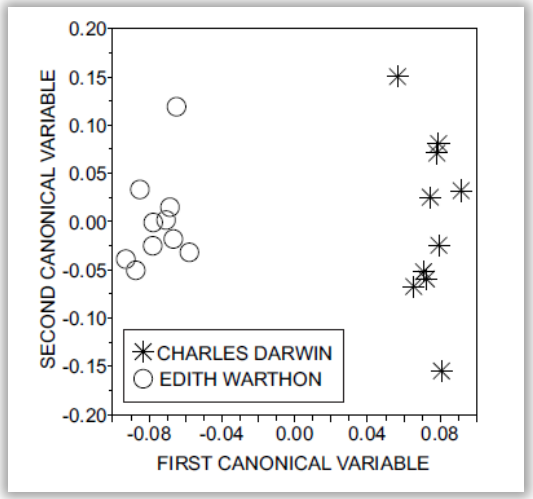
Literary Movements

Relationship between the best clustering of writing styles the traditional classification of literary movements.

Cluster Boundary	Literary Boundary	Literary Movement
1590 – 1653	1558 – 1903	Elizabethan era
1664 – 1761	1660 – 1798	Neoclassicism/ Enlightenment
1767 – 1793	1660 – 1798	Neoclassicism/ Enlightenment
1794 – 1818	1764 – 1820	Gothic fiction
1826 – 1906	1830 – 1900	Realism
1826 – 1906	1865 – 1900	Naturalism
1906 - 1922	1890 - 1940	Modernism



Identification of movements using complex networks to represent text

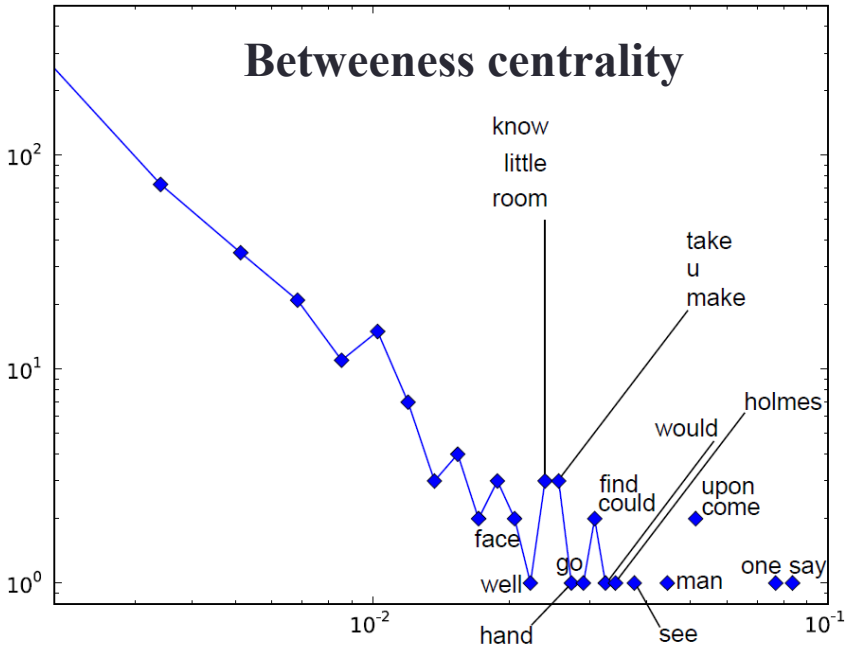


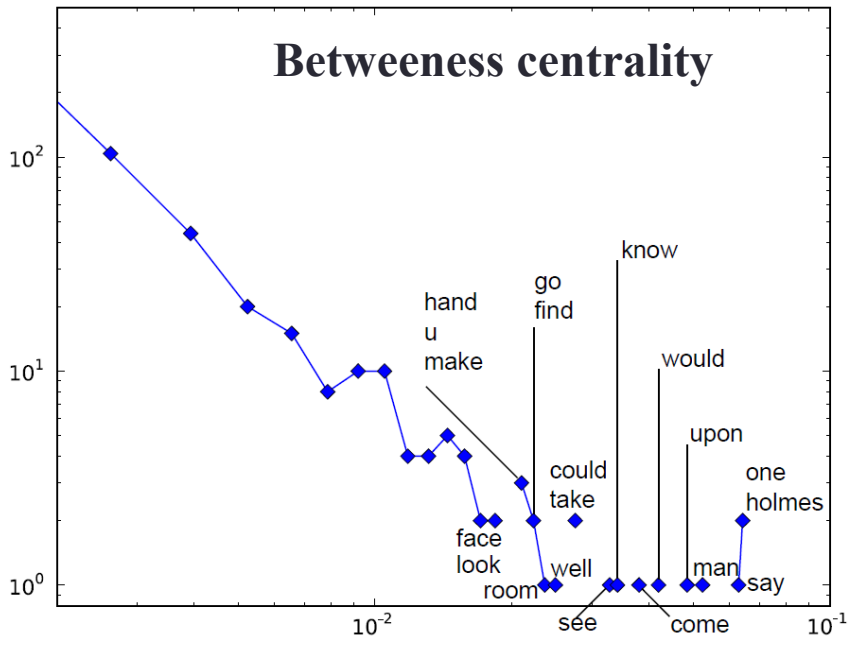
Darwin vs Edith Warthon

Amancio et al, *New J. Phys. 2012*

Role of words in network structure

Dissimilarity matrices for 4 metrics (degree, shortest paths, betweeness, intermittency). Only for 100 most relevant nodes



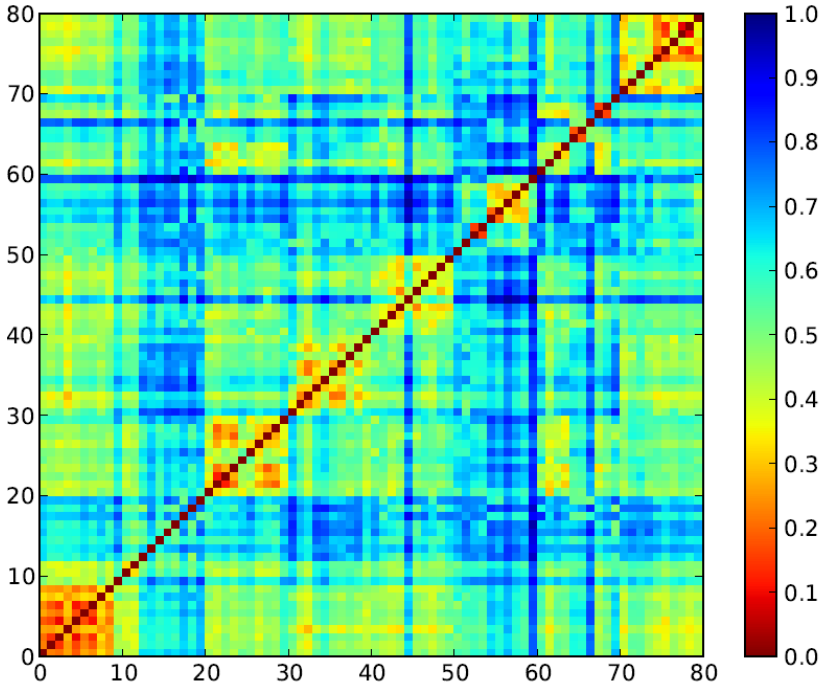


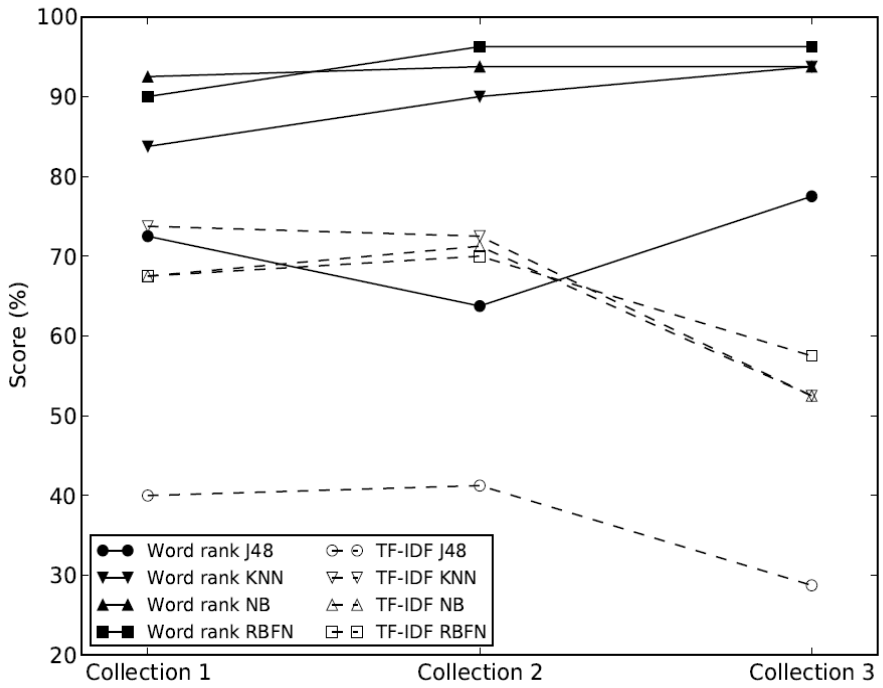
Two Sherlock Holmes novels: Only one different word among the first 20

Similarity (dot product) between 2 texts is high if the same words occupy similar positions in the distributions

Akimushkin et al., Physica A., 2018

Efficient authorship attribution



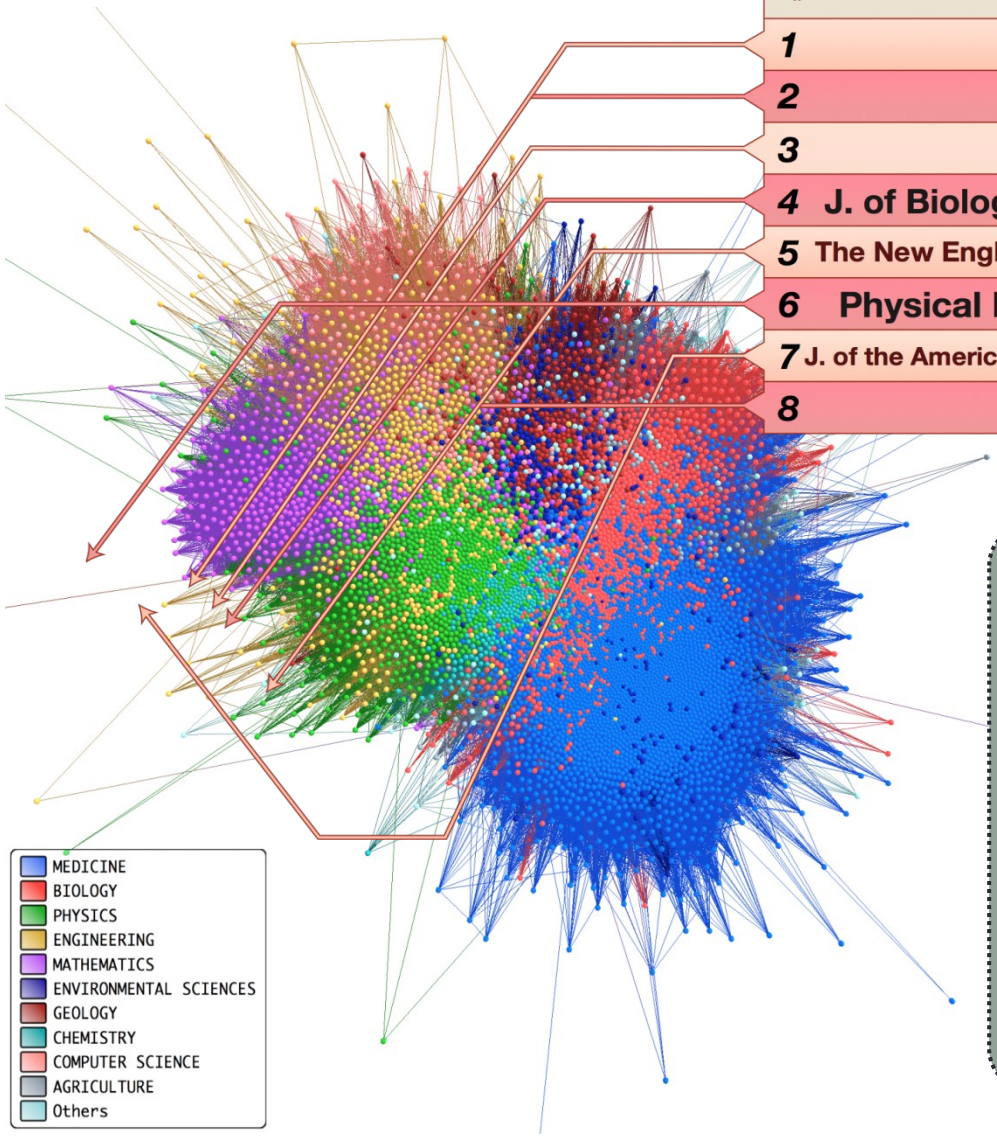


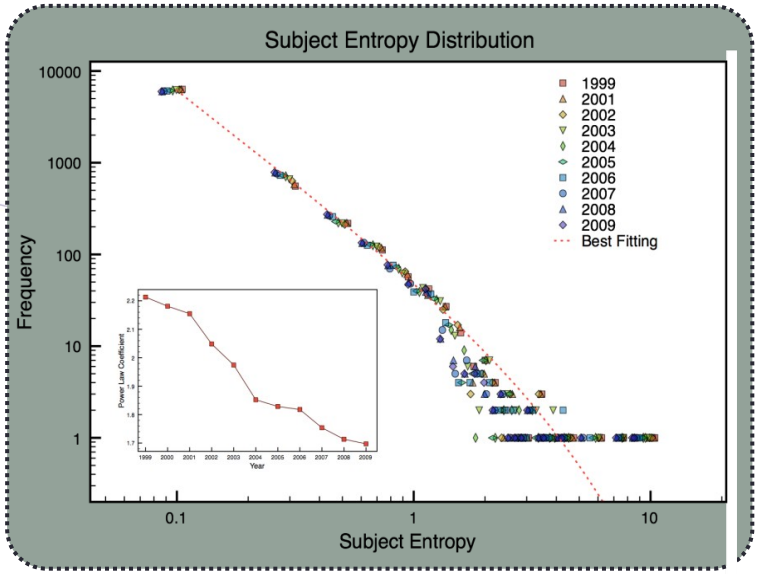
Dissimilarity matrices with multi-dimensional scaling for feature selection. Radial Basis Function Network (RBFN) yields the highest scores

Akimushkin et al., Physica A., 2018

Multidisciplinarity in Science

#	Journal	Subject Entropy
1	Nature	8.61
2	Science	8.38
3	PNAS	7.22
4	J. of Biological Chemistry	5.07
5	The New England J. of Medicine	4.20
6	Physical Review Letters	4.08
7	J. of the American Chemical Society	3.52
8	The Lancet	3.43





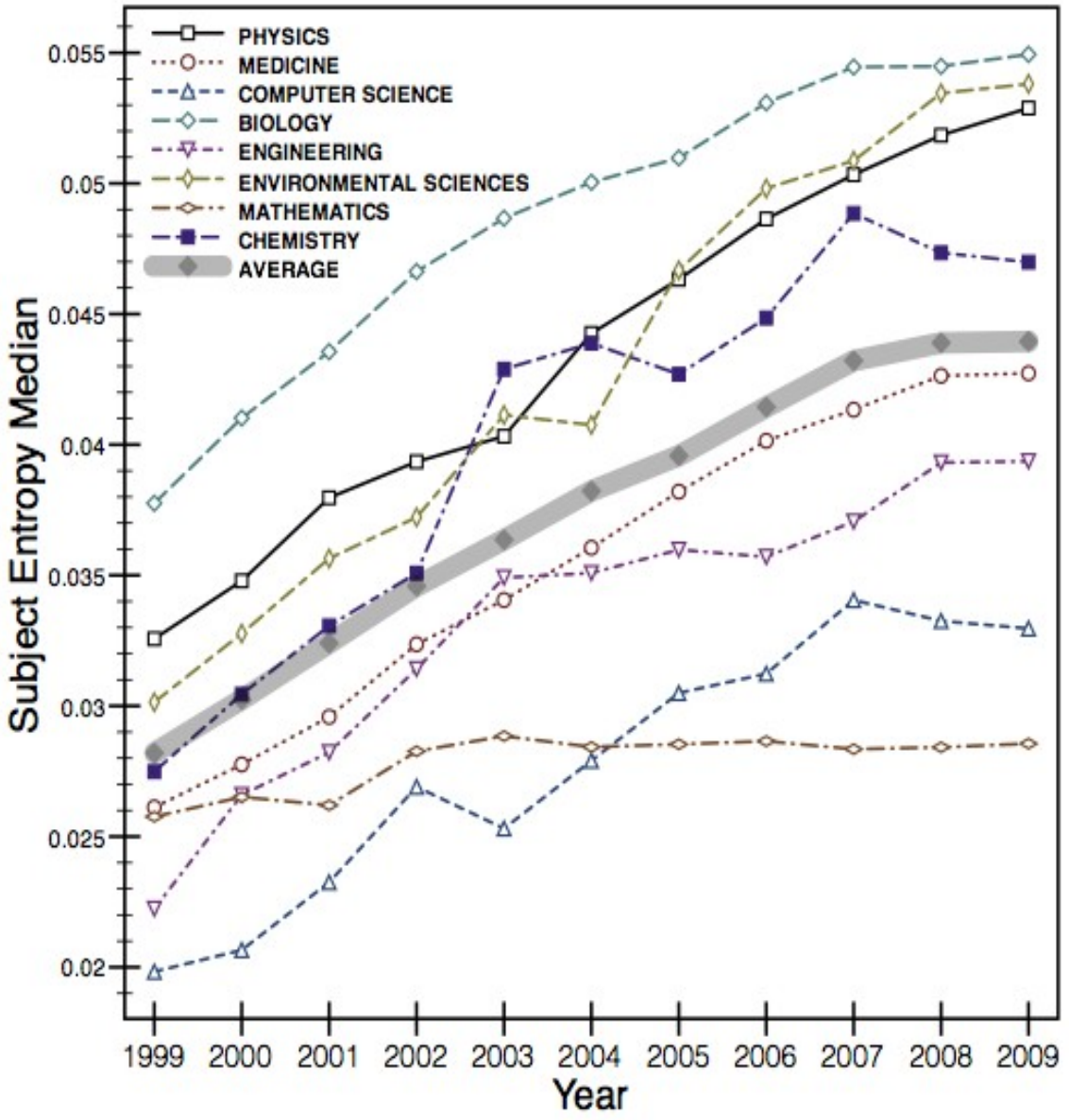
Filipi N. Silva et al., J. Informetrics, 2013

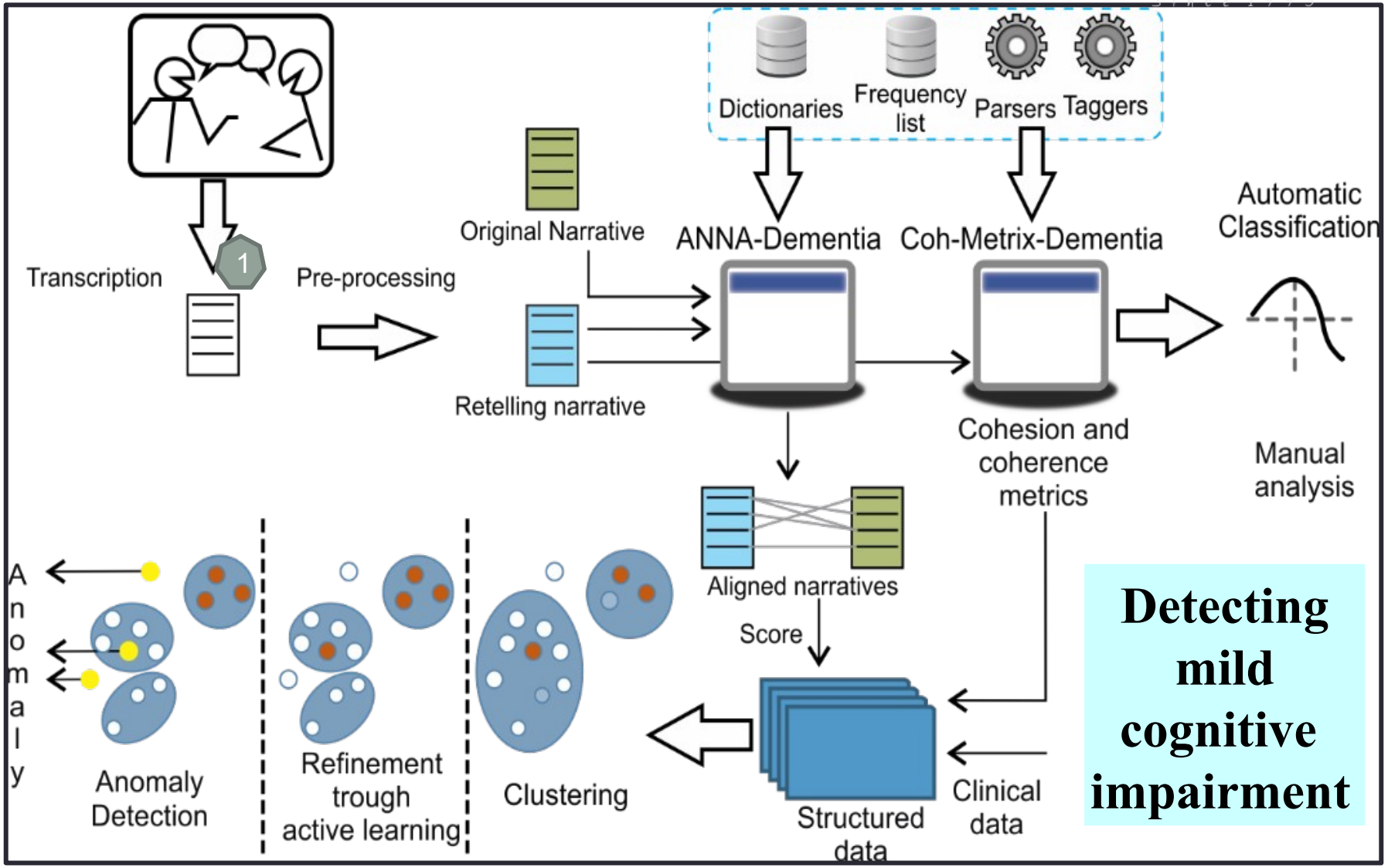
Multidisciplinarity in Science

- Areas becoming increasingly multidisciplinary

- Subject entropy correlates highly with impact factor and betweenness centrality

Filipi N. Silva et al.,
J. Informetrics, 2013





With Sandra Aluísio, Letícia Mansur and Diego Amancio, ACL 2017

Panorama da Pesquisa em PLN para o Português

Objetivo

- Panorama de PLN para a língua portuguesa.

Metodologia

- Busca no OpenAlex com os termos: “PLN”; “NLP”; “processamento de linguagem natural”; “natural language processing”
- Total recuperado: **3.303 artigos**.
- Rede final: **1.766 artigos** na maior componente conectada.
- Similaridade calculada com embeddings do modelo multilíngue **SentenceTransformer**.
- Limiar de similaridade: **70%**.

Principais desafios

- Baixa visibilidade internacional de trabalhos em português (< 5000 artigos na Web of Science or OpenAlex).
- Fóruns importantes (ex.: PROPOR) nem sempre indexados.
- Ambiguidade entre IA, aprendizado de máquina e PLN.
- Necessidade de modelos multilíngues para textos em português.
- Possibilidade de mapear tópicos emergentes e relações entre subáreas.

Rede de similaridade

OpenAlex: “PLN” OR “NLP” OR “processamento de linguagem natural” OR “natural language processing” - 3.303 artigos; 1.766 na maior componente conectada. Limiar de similaridade 70%.

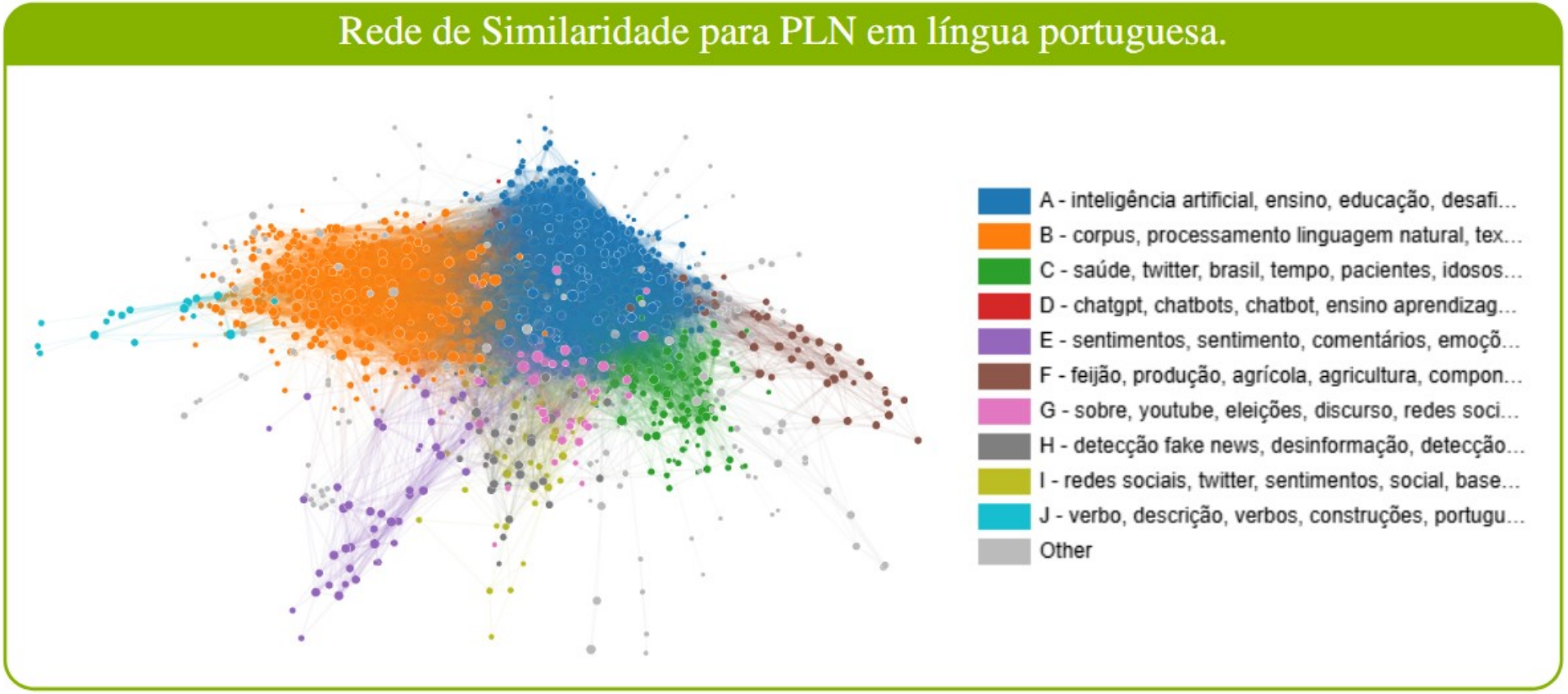
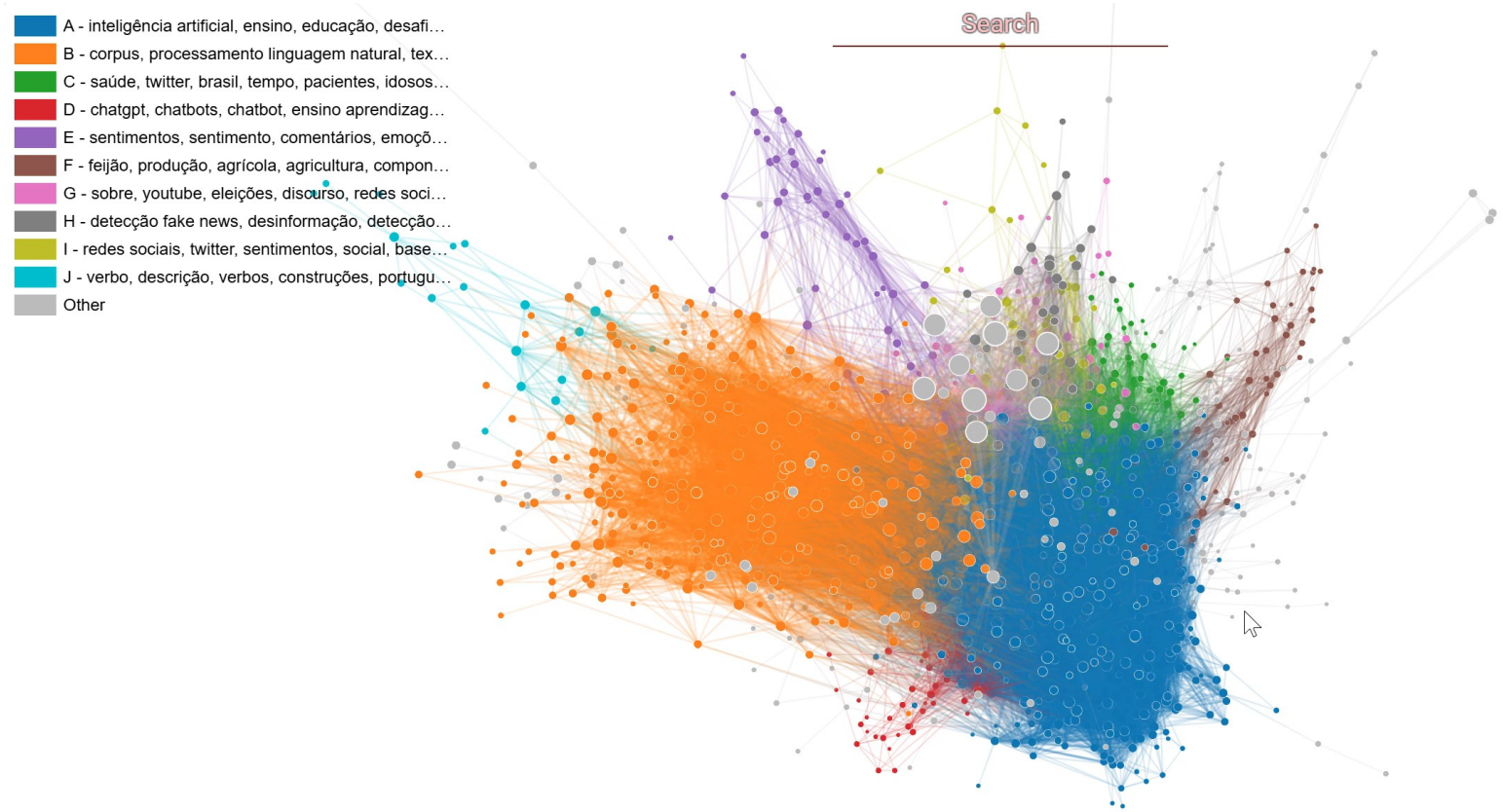


Figura 6: Rede de similaridade para publicações sobre Processamento de Linguagem Natural em português, mostrando as principais comunidades temáticas e suas inter-relações.


A - inteligência artificial, ensino, educação, desafi…
B - corpus, processamento linguagem natural, tex…
C - saúde, twitter, brasil, tempo, pacientes, idosos…
D - chatgpt, chatbots, chatbot, ensino aprendizag…
E - sentimentos, sentimento, comentários, emoçõ…
F - feijão, produção, agrícola, agricultura, compon…
G - sobre, youtube, eleições, discurso, redes soci…
H - detecção fake news, desinformação, detecção…
I - redes sociais, twitter, sentimentos, social, base…
J - verbo, descrição, verbos, construções, portugu…
Other
Search
press space to start the layout
SVG
Size
Color
ClusterName
Category10
Edges

Principais Comunidades Temáticas

Comunidade A (748 artigos)

- Inteligência artificial aplicada à educação, ensino e tecnologias.
- Muitos artigos ligados à IA em geral.
- Muitos artigos utilizam processamento de texto, mas nem todos são PLN estrito.

Comunidade B (377 artigos)

- Núcleo principal de PLN.
- Construção de corpora.
- Classificação automática de textos.
- Linguística computacional e modelos linguísticos para português brasileiro.

Outras comunidades relevantes

- **C:** aplicações em saúde e COVID-19.
- **D:** chatbots e ChatGPT.
- **E e I:** análise de sentimentos e Twitter.
- **G:** eleições, política e redes sociais.
- **H:** fake news e desinformação.
- **J:** estudos linguísticos e análise semântico-discursiva.

Interpretação

- As comunidades A e B dominam a estrutura da rede.
- A comunidade B representa os grupos mais diretamente ligados ao PLN.
- As demais comunidades conectam PLN a aplicações em IA.

Número de nós, palavras-chave e descrição de cada comunidade.

Comunidade	Nós	Palavras-chaves	Descrição
A	748	inteligência artificial, ensino, educação, desafios, aprendizado, tecnologias, científica, máquina, papel, ia	Aplicações de inteligência artificial no contexto educacional; tecnologias e estratégias para ensino, aprendizagem e desenvolvimento científico.
B	377	*corpus*, processamento linguagem natural, textos, automática, classificação, linguística, português brasileiro, textual, língua, modelos	Processamento de linguagem natural e linguística computacional; análise automática de textos, classificação textual, *corpora* e modelos linguísticos em português brasileiro.
C	110	saúde, twitter, brasil, tempo, pacientes, idosos, in, pandemia covid-19, tratamento, revisão	Uso de redes sociais para monitoramento de saúde pública; análise da pandemia de COVID-19, impacto em pacientes e idosos, e revisão de tratamentos no Brasil.
D	53	chatgpt, chatbots, chatbot, ensino-aprendizagem, sobre, revisão, ferramenta, auxiliar, assistente, sistemática	Desenvolvimento e aplicação de chatbots e assistentes virtuais; ferramentas para ensino e aprendizagem e revisões sistemáticas de tecnologias educacionais.
E	47	sentimentos, sentimento, comentários, emoções, português, métricas, mídias sociais, mineração, rede, mapeamento	Análise de sentimentos em mídias sociais; mineração de dados textuais, métricas emocionais e mapeamento de interações em português.
F	46	feijão, produção, agrícola, agricultura, componentes, rendimento, produtividade, cultura soja, seleção, gesso	Produção e produtividade agrícola; técnicas de cultivo, seleção de componentes e otimização de rendimento em culturas como feijão e soja.
G	41	sobre, youtube, eleições, discurso, redes sociais, deputados, discursos, ódio, sentimento, twitter	Análise de discursos e sentimentos em redes sociais durante eleições; estudo de conteúdo em YouTube e Twitter, incluindo discursos de deputados e manifestações de ódio.
H	38	detecção fake news, desinformação, detecção automática, notícias falsas, jornalismo, utilizando, produção, covid-19 brasil, language, avaliação	Detecção automática de fake news e desinformação; análise de notícias falsas e produção jornalística durante a pandemia de COVID-19 no Brasil.
I	37	redes sociais, twitter, sentimentos, social, baseado, tweet sobre, eventos, classificação, identificação, caso	Análise de sentimentos e classificação de eventos em redes sociais; identificação de padrões sociais e comportamentais no Twitter.
J	18	verbo, descrição, verbos, construções, português, linguagem natural, processamento, base, verbal, semântico-discursiva	Estudos linguísticos em português; análise de verbos, construções verbais e descrições semântico-discursivas utilizando processamento de linguagem natural.

Comentários Finais

- Redes complexas são úteis para várias aplicações de PLN
- Precisamos de redes complexas se temos os LLMs?
- A análise de comunidades e métricas topológicas permite identificar padrões semânticos e relações estruturais relevantes em grandes volumes de dados.
- Quanto de PLN (tradicional) ainda podemos empregar com os LLMs?

Agradecimentos

Graça Nunes e Helena Caseli

Brasileiras em PLN