TEXATA

Big Data Analytics Showdown 2015

Auto-answering in the semantic space

Álvaro Barbero Jiménez

Chief Data Scientist

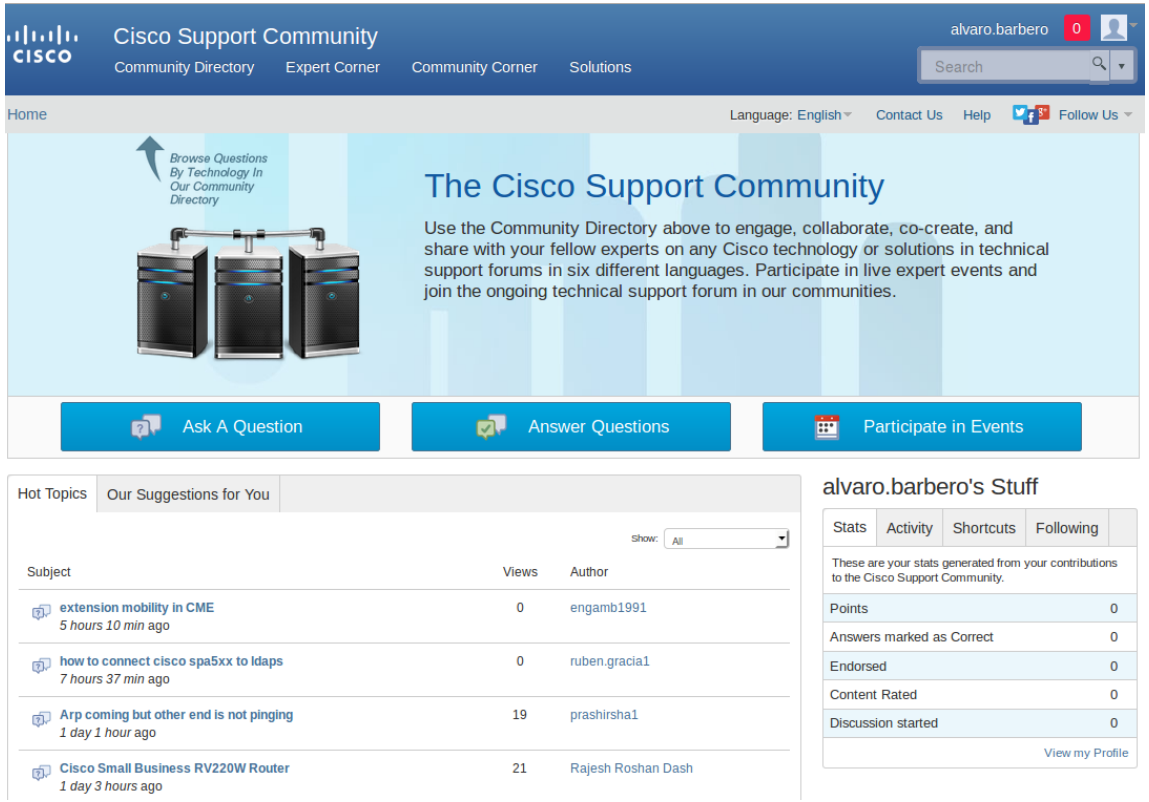
instituto de ingeniería del conocimiento

www.iic.uam.es

8 de noviembre de 2015

Customer service doesn't scale

Even with the use of online boards and communities!



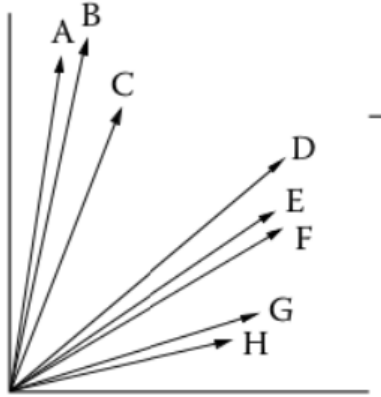
X Users repeatedly ask the same or very similar questions

Q: how can we improve the speed of question answering?

What stackoverflow did

…can we automate this?

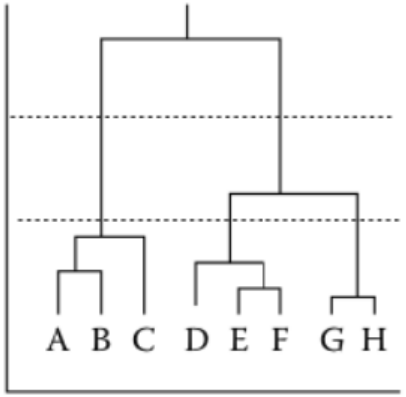
The key idea: semantic embeddings



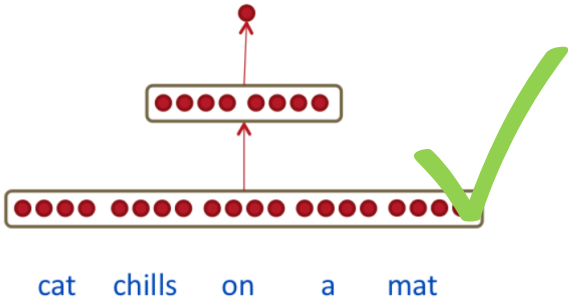
Instances of polysemous words in context (vectors)

	A	B	C	D	E
A	0	2	4	13	17
B	2	0	3	11	14
C	4	3	0	9	12
D	13	11	9	0	2
E	17	14	12	2	0

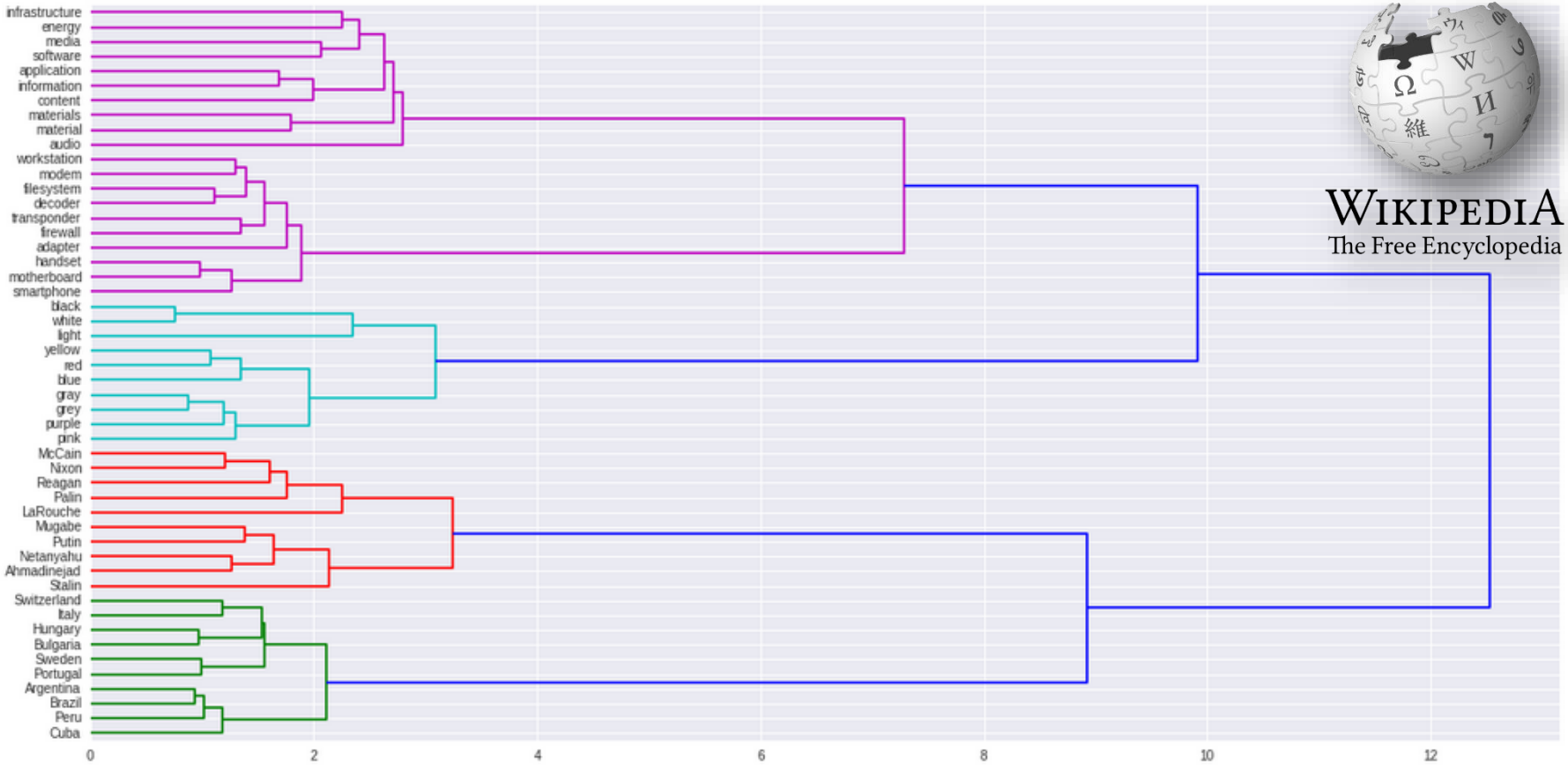
Derived distances and possible singular value decomposition, then clustering



Derived clusters (and possible cuts)

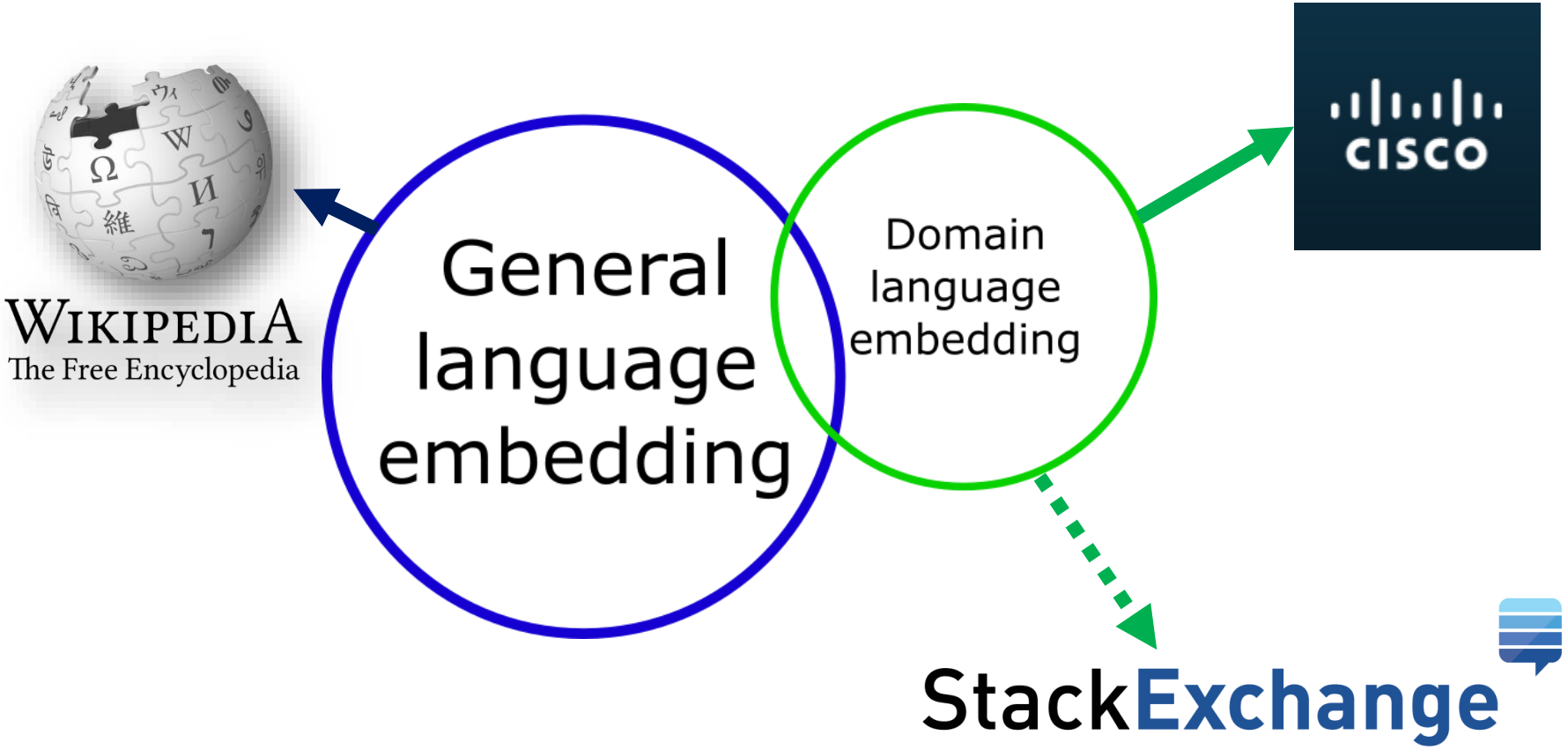


English Wikipedia semantic embedding



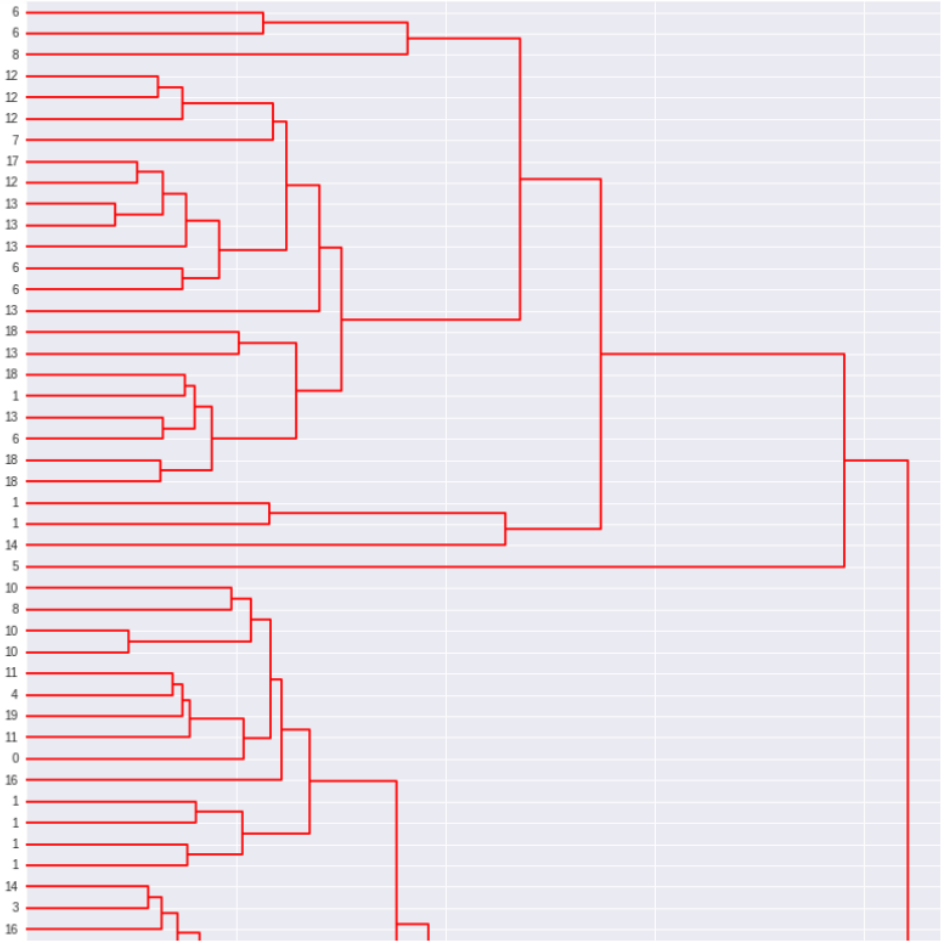
Polyglot embeddings: https://sites.google.com/site/rmyeid/projects/polyglot#TOC-Download-the-Embeddings

Combining embeddings



CISCO posts embedding

Post ID



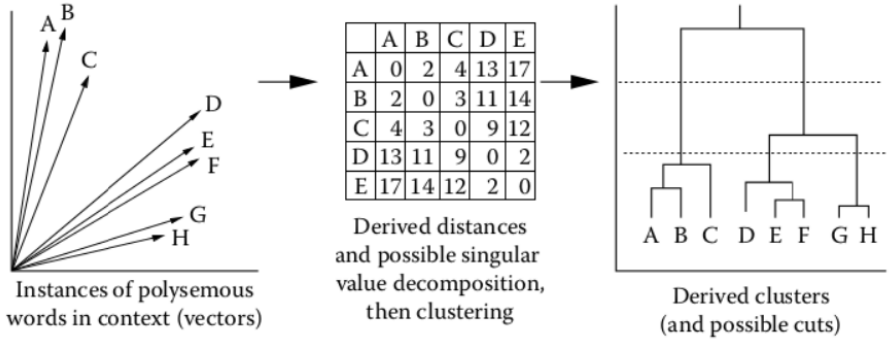
Label agreement score on StackExchange community

$$score = \left(\sum_{tags(d_i) \cap tags(d_j) \neq \emptyset} simil(d_i, d_j) \right) - \left(\sum_{tags(d_i) \cap tags(d_j) = \emptyset} simil(d_i, d_j) \right)$$

WIKIPEDIA The Free Encyclopedia	WIKIPEDIA The Free Encyclopedia
0.519	0.730

Fast search filtering

~100



	A	B	C	D	E
A	0	2	4	13	17
B	2	0	3	11	14
C	4	3	0	9	12
D	13	11	9	0	2
E	17	14	12	2	0

- Fast TF-IDF search
- Horizontally scalable

- Fine-detail search for semantics

Conclusions

- ✓ Real-time recommendations for already existing solutions to new questions might be possible using semantic embeddings

  - ✓ Faster troubleshooting for users
    → Increase in customer satisfaction

  - ✓ Reduction of repetitive burden on technical support team
    → Increase in productivity

C/ Francisco Tomás y Valiente, nº 11
EPS, Edificio B, 5ª planta
UAM Cantoblanco. 28049 Madrid
Tel.: (+34) 91 497 2323

www.iic.uam.es
blog.iic.uam.es

Elementos gráficos de apoyo obtenidos en:
designed by freepik.com
pixabay