Predicting Wikipedia Article Translation Counts

by Jay Park

EN ∨

OUTLINE

Contents [hide]

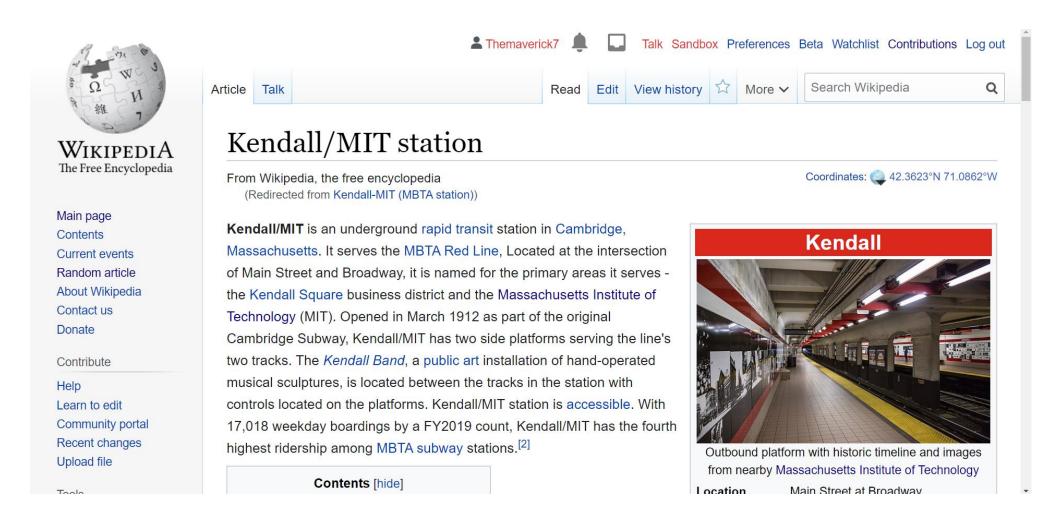# Wikipedia is the 13th most visited website worldwide

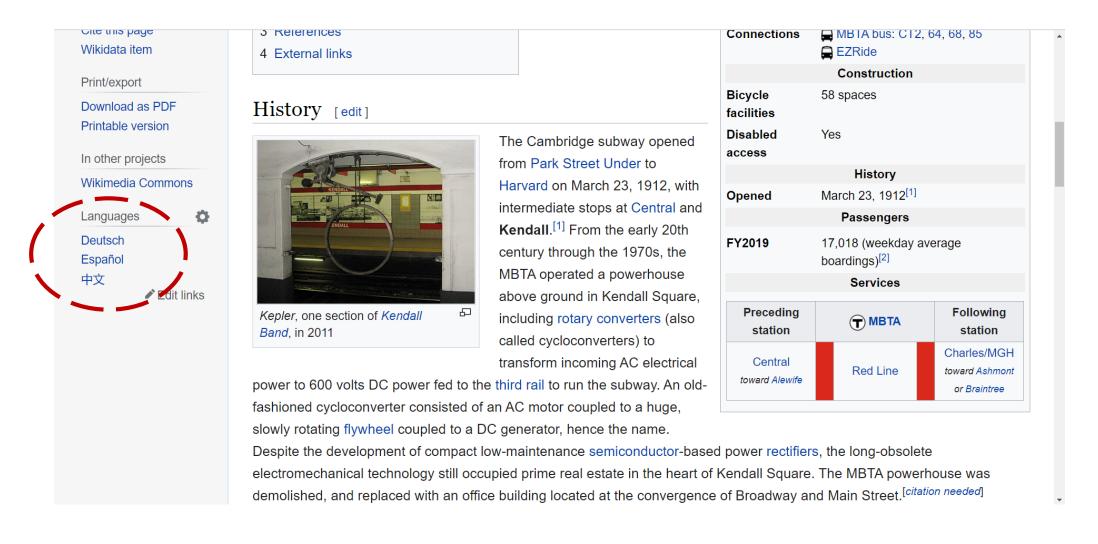# Wikipedia is the 13ᵀᴴ most visited website worldwide

- Founded in 2001
- Available in 285 languages
- The English Language Wikipedia is the largest
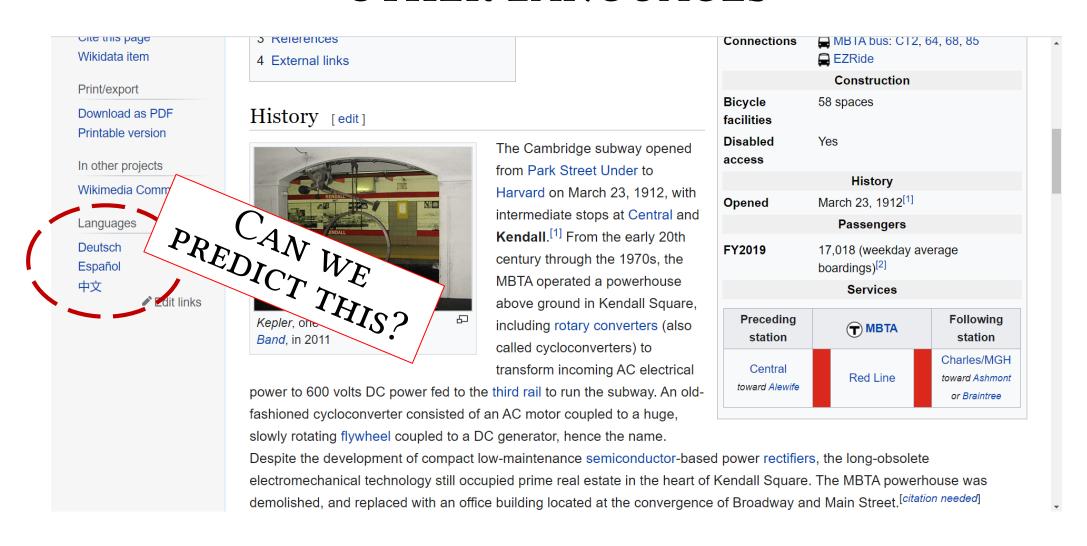- The official song of Wikipedia is "Hotel Wikipedia"

ARTICLES ARE OFTEN TRANSLATED/WRITTEN IN OTHER LANGUAGES

Articles are often translated/written in other languages
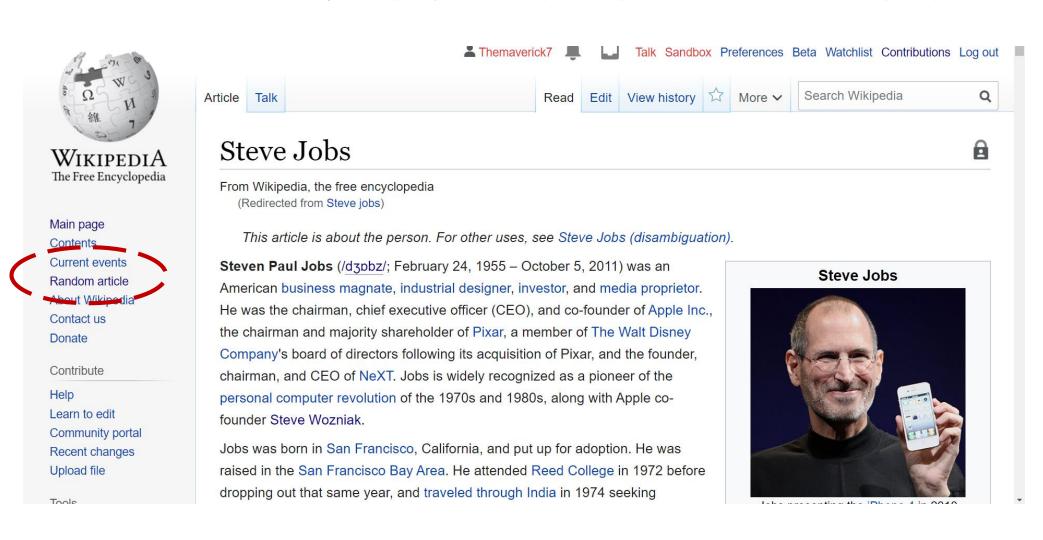
Articles are often translated/written in other languages

Outline

Contents [hide]

3655 Articles were scraped using the 'Random article' feature

Selected features for regression

- Body text length

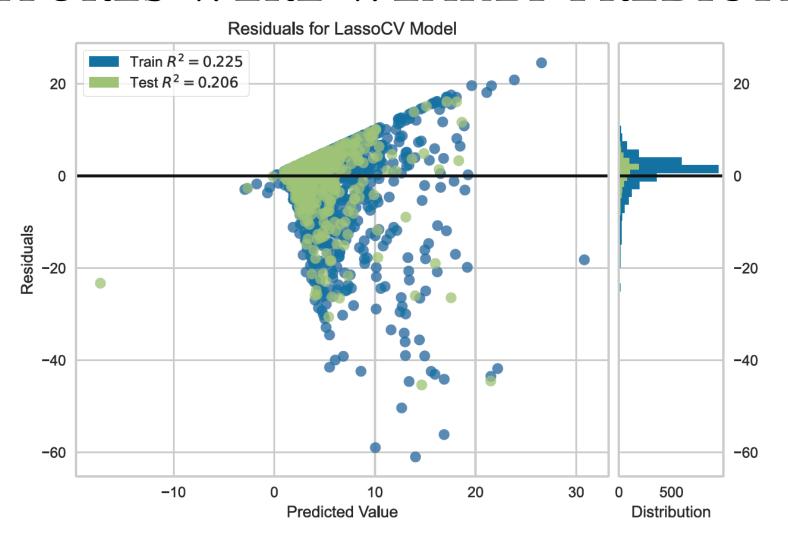- Days since article creation

- # of references

Selected features for regression

- Body text length

- Days since article creation

- # of references

- Total length of text in tables

- # of "Main article" links

- # of "See also" links

- # of "Further information" links

- # of links to other wiki articles

- Is sports?

REGRESSION USING MANUALLY-CURATED FEATURES WERE WEAKLY PREDICTIVE

Residuals for LassoCV Model

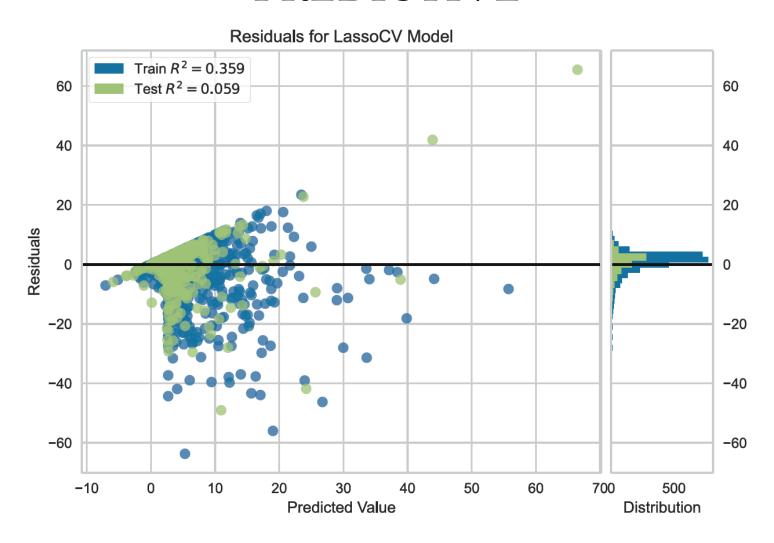Train R^2 = 0.225
Test R^2 = 0.206

NLP method: Bag of Words

This is a article about the future of our country. It looks quite grim. However, we believe we can make it through all the hardships like we have before. Our country is strong. The author of this article approves this message.
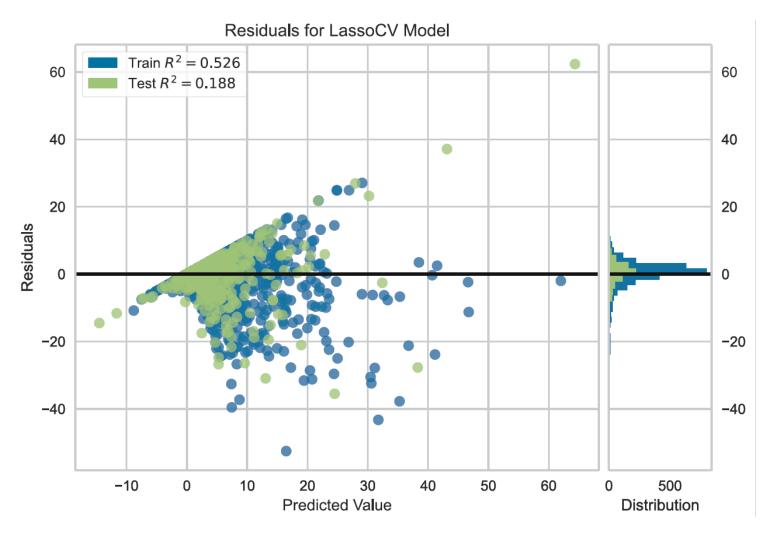
Cleanup

'article'	'future'	'country'	'comfort'	'author'	'message'	
2	1	2	0	1	1	. . .

Regression using basic NLP was less predictive

COMBINATION OF THE TWO APPROACHES DID NOT YIELD AN NOTICEABLE IMPROVEMENT

Residuals for LassoCV Model

OUTLINE

Contents [hide]

CONCLUSIONS AND FUTURE DIRECTIONS

- # of languages can be weakly predicted

- More refined NLP methods, or neural networks

- Try incorporating PageRank

THANK YOU WIKIPEDIA!