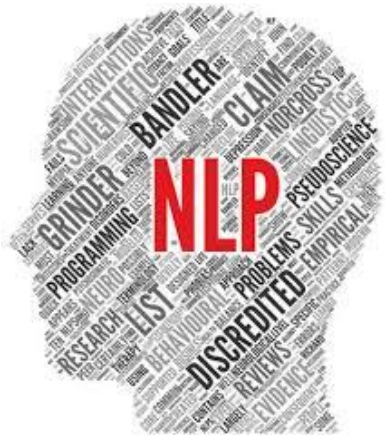
How to show related content using NLP in Drupal 8?



open sense labs

Drupal™

It's nice and customary to begin with definition.

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken.

1. “Hey Siri. Can you find a place to party near DrupalCamp Atlanta venue?”.

   Siri : “No you have to understand NLP first.”

1. “Ok Google!. Can you find flights from Atlanta to Delhi?”.

   “Google : I don’t think you should go back.”

What actually happens.

1. Analyze
2. Understand
3. Derive meaning from human language.
4. Reply in smart and useful way.

What can NLP do?

1. Information Extraction
2. Summarization
3. Text Classification
4. Speech Recognition
5. Question Answering

Algorithms

We actually use this term when we don’t want to explain what’s actually happening.

1. Probabilistic language models based on n-grams recover a surprising amount of information about a language. They can perform well on such diverse tasks as language, identification, spelling correction, genre classification, and named-entity recognition.
2. Text classification can be done with naive Bayes n-gram models or with any of the classification algorithms we have previously discussed. Classification can also be seen as a problem in data compression.

Continued…(Just a lil more.)

3. Information retrieval systems use a very simple language model based on bags of words, yet still manage to perform well in terms of recall and precision on very large corpora of text. On Web corpora, link-analysis algorithms improve performance.

4. Information-extraction systems use a more complex model that includes limited notions of syntax and semantics in the form of templates. They can be built from finite state automata, HMMs, or conditional random fields, and can be learned from examples.

Problems we are facing...

Although they never end, let’s just make a list of few.

1. Tagging content.
2. Summarizing content.
3. Content Duplicacy.

How big is the problem?

1. The Post publishes an average of 1,200 stories, graphics, and videos per day.
2. At 15 million products, Amazon India has highest number of products.
3. “NYTimes.com publishes roughly 150 articles a day (Monday-Saturday), 250 articles on Sunday and 65 blog posts per day,”
4. Google now processes over 40,000 search queries every second on average (visualize them here), which translates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide.Till when they will be able to provide nice recommendation.

How these are solved in a Drupal website?

Traditional methods.

1. Content with similar tags is used to show related content?

Disadvantages : Not possible with large content, tags duplicacy, difficult to remember all the tags, time needed for tagging.

2. Use first few words/characters as a summary.

NLP for auto-tagging and related content.

Body Field -> NLP Engine -> Tf-iDf -> Important Words -> Add to Tag Field -> Modify -> Related Block.

Auto Summary

Body Field -> Lex Ranking -> Summary.