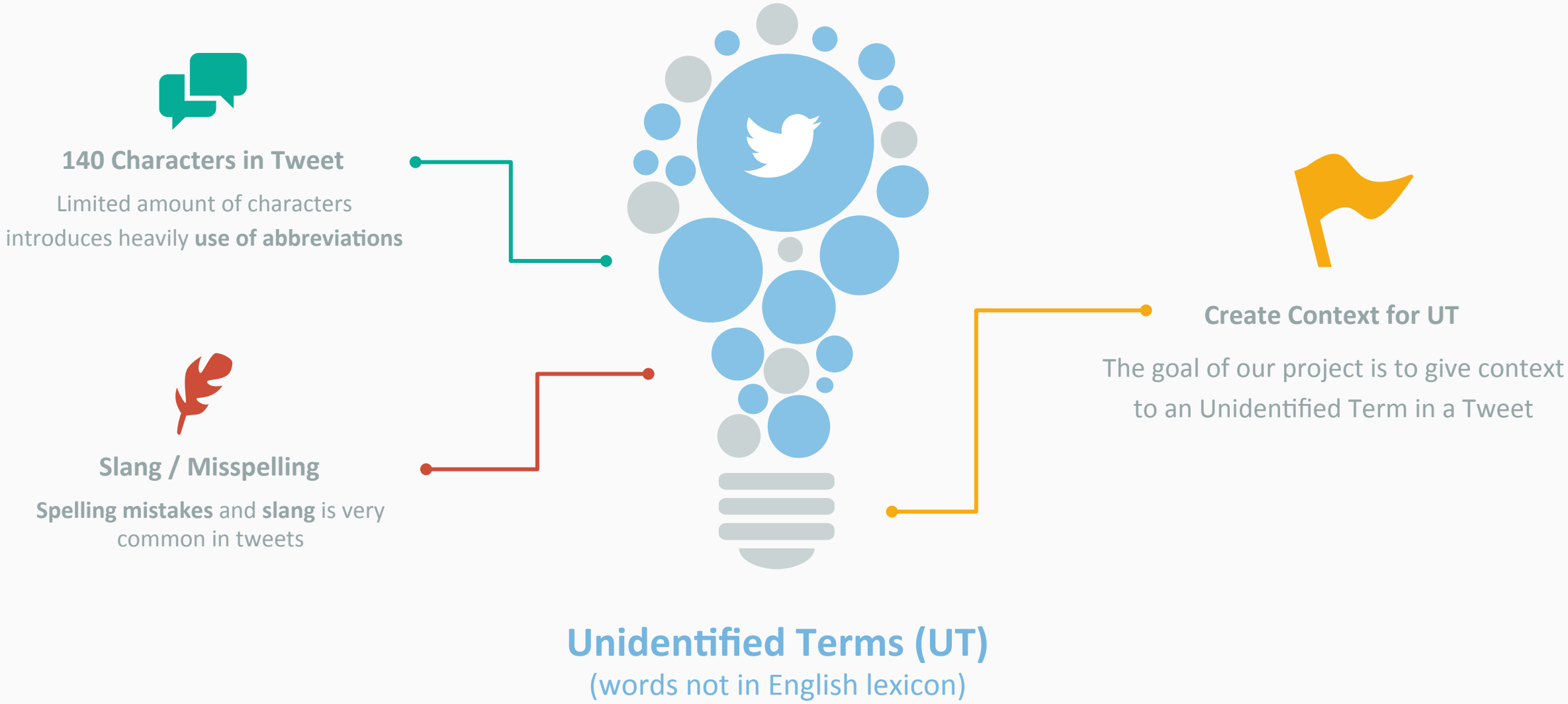


Determine the Meaning of Unknown Terms in Tweets




THE PROBLEM

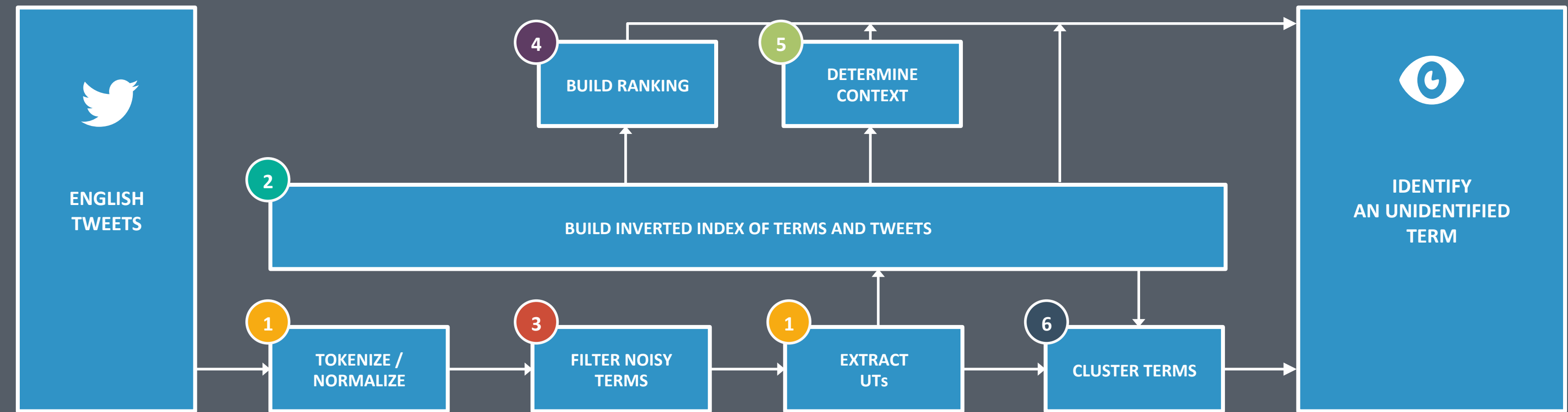





PROJECT GOAL



USED ARCHITECTURE





INDIVIDUAL SUBTASKS



TOKENIZE / NORMALIZE & EXTRACT
Unidentified terms are found by taking the list of all tweets and filtering away all known words. To catch all known words, we apply stemming and lemmatization to match as much words as possible.

1

BUILD INVERTED INDEX OF TWEETS & TERMS

- Tweets - using a document-matrix
- Terms - using a term-by-term matrix with support for retrieving: term frequency, postings lists and TFIDF

2

FILTER NOISY TERMS

Noise: Emoticons, Terms with special characters, random keystrokes, etc...
Filtering with set of regular expressions intersected with IDF-score of term above threshold

3

BUILD RANKING

Using normalized Tweets metrics (#retweets, #favorites, #author_followers) to calculate score of tweet for a UT and return top X

4

DETERMINE CONTEXT

Association rule mining using the Apriori algorithm to determine what terms determine the overall context co-occurring with the unidentified term

5

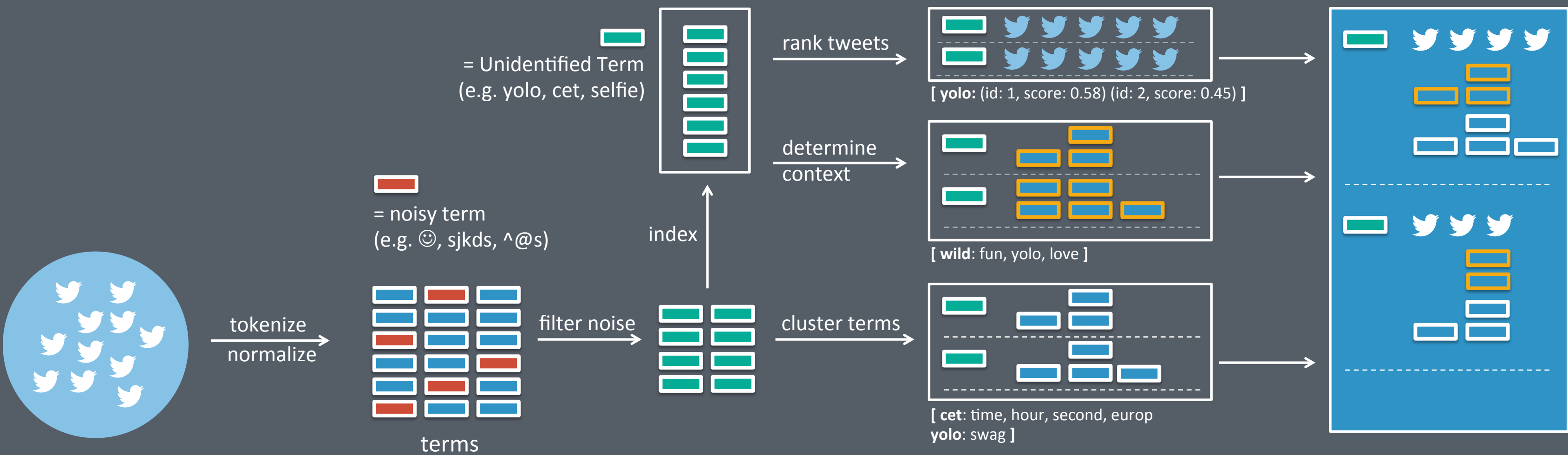
CLUSTER TERMS

Co-occurring words are found by taking the intersection between the tweet terms and the UTs. For every UT we then record all the words in the tweet as co-occurring words. Afterwards, these words are sorted per UT and words that occur below a threshold are thrown away.

6



EXAMPLE USE CASE



CONCLUSIONS

- Useful information found on Unrecognized Terms
- Not able to determine the **actual** meaning of a term



RECOMMENDATIONS

- Increase initial tweet collection
- Use of external corpora and wordbases

