Haikus are decent
As a summary of news
But serve more as art

Emma Dickson
Courant Institute
New York Institute
ejd315@nyu.edu

Abstract

My final project is a chrome browser extension that creates a haiku from the text of the visited page. It's designed to be used with news articles and aims to create haikus that function both as a useful summarization of the text and amusing poems. Ultimately though the algorithm used correctly picked out key words the limitations of the haiku structure made the resulting poems more useful as amusements than as summaries.

1 Introduction

Using javascript the text on the page is encoded as utf-8 and sent to a python flask app backend when the plugin icon is pressed. Once in the python flask app the text has some extraneous characters removed. The text is then shingled into three and two word overlapping segments. At this stage helping dictionaries are created, two for each shingle size. One of which uses POS sequences as keys and maps to shingles while the other uses the shingle and maps to POS sequences. As I'll discuss below the lines of the haiku are constructed from shingles of the article that receive the highest TFIDF score in comparison to the corpus. The dictionaries created in the pipeline aid in making these constructions grammatically correct and ensuring that the poem has an overall coherent structure.

2 The Approach

2.1 How to correctly count syllables

The cmu pronunciation dictionary is an excellent resource for those who want to count syllables. Available in the nltk package you can easily and quickly parse the number of syllables in any given word. Of course dictionaries cannot be all things to all people and a strategy for handling OOV words must be set in place. For the purposes of this project I used a fairly simple yet accurate algorithm. Given an OOV word, I counted the number of vowels and then added or subtracted to this count based on a list of specialty rules. I established two lists of syllable exceptions, for example “ily” contains two vowels yet is one syllable so a word that contains ily would have 1 subtracted from its total syllable count. I based my initial rules list off of the algorithm used by syllablecount.com and refined upon their basic lists through trial and error which I found to be fairly reliable. The rules employed can be found in the Constants.py file

2.2 How to best rank words

I used a tfidf matrix to rank potential shingles in comparison to my corpus. Each word in a shingle was considered and added to form a total tfidf score for the shingle. Shingles were then placed in a list and sorted so by tfidf score so that the highest one could be easily chosen.

2.3 Corpus Selection

My corpus consisted of 55 individual documents. 50 from the Reuters training corpus included in the nltk package and 5 from the brown corpus across various news/informational genres. This was done to maximize the plugins effectiveness upon news articles. In a more advanced implementation persistent storage could be implemented and the text on sites visited continually added to the corpus.

2.4 Implementing Grammar Rules

I used a series of constant grammar rules to maximize the value of these haikus as summaries and poems. Grammar carries over line by line and there are a set list of POS tags acceptable for haikus to start and end with. POS tags are attributed to the shingles using the nltk package early on in the preprocessing stage and stored in a series of dictionaries that can be referenced during the line generation.

3. Experiments

Dataset: Two participants were asked to evaluate 20 haikus generated across five different news platforms in terms of their effectiveness as a summary as well as their merits as a poem. The sites used were not present in the training corpus and the articles used varied both in length and the genre of news. Though it was not evaluated separately it was noted by participants that the plugin performed better on international and "issue" news in comparison to op-eds, pieces focusing on a single individual or sports news. This can likely be attributed to the training corpus and could be tweaked in the future. The results of the evaluation can be seen in the attached table and are discussed at length below.

	Summary Evaluations			
	Participant #1	Participant #2	Overlap	Actual
Good	2	3	2	2
Neutral	11	11	11	8
Bad	7	6	6	4
	Poem Evaluations			
	Participant #1	Participant #2	Overlap	Actual
Good	10	7	7	6
Neutral	7	9	7	5
Bad	3	4	3	2

Table 1

Baseline: Since this project was an original idea the baseline is simply my earlier implementations. Originally I had liked the idea of choosing a random number and selecting the first word of each line based on words that fit the appropriate pos tags and syllable count based off the random number. This approach while more effective resulted in poems that didn't flow well or stand on their own outside of summarization purposes.

Results: Participants overlap in their evaluation of the haiku's effectiveness as a summary and a poem resulted in the following expected averages and observed odds:

Expected Average Odds of Agreement		
	Summarization	Poem Quality
Good	1.25	42.5
Neutral	55	40
Bad	32.5	17.5
Observed Odds of Agreement		
	Summarization	Poem Quality
Good	1	30
Neutral	40	25
Bad	20	10

Table 2

(All numbers represent percentages)

Using Cohen's formula for calculating the kappa I got the following scores:

Kappas Scores		
	Summarization	Poem Quality
Good	0.0025	-0.22
Neutral	-0.33	-0.25
Bad	-0.19	-0.09

Table 3

In addition the averages of all the scores were calculated and can be seen below.

	Summarization Scores		Poem Merits Scores	
	Participant 1	Participant 2	Participant 1	Participant 2
	2	2	3	2
	2	2	3	3
	2	2	3	3
	1	2	2	3
	2	2	3	3
	2	2	3	1
	1	1	2	1
	2	2	2	2
	2	1	3	3
	2	2	2	2
	1	1	3	2
	2	1	1	1
	1	1	1	1
	2	2	2	2
	1	2	2	2
	1	1	2	2
	3	3	3	3
	2	3	3	2
	3	3	3	3
	1	2	1	2
Average Score	1.75	1.85	2.35	2.15

Table 4

The implications of these findings is discussed in the conclusions section of this paper. Some raw data has been included in the following section to allow the reader to get a sense of the types of poems produced.

Data: A list of example poems followed by their corresponding links

1. bolton mccain

   graham recognized in

   john bolton bolton

2. trump or mcmahon
   the audience pays
   at trump properties

3. the fake stories
   hack election systems
   believe the fake

4. gawker vs gamergate
   refreshing gawker pages
   are sleeping giants

5. the costumes teams
   batman spiderman gumby
   prepares to leave

6. The election the
   intelligence mr putin
   that the computer

7. The time comey
   traveling the united
   comey and mueller

8. allepo's rebel
   remaining rebel areas
   leave your friends

9. trump winning donald
   immigration and trump
   is sessions trump

10. the broadcasters could
    follow homepage news
    winning donald trump

11. alan thicke was
    pop culture aspiring
    canada alan

12. read on jezebel
    was humiliating his
    eligible read

13. read on jezebel
    was never recovered
    read on fusion

14. women the texas
    texas regulation
    says kristi hamrick

15. policy the trump
    lifting western sanctions
    spotted in moscow

16. month 's election
    did interfere senate
    cia 's intelligence

17. trump or mcmahon
    the ring advertisement
    returned to trump

18. the abuse nor
    serial sexual abuse
    appeared the chants

19. kesha accepted the
    friday kesha accepted
    slate staff writer

20. christmas on hip-hop
    invoke well-known christmas
    is disgusting read

21. flavors the untold
    story of american
    washington lohman

22. the charleston post
    including nine murder
    testified that roof

23. bank 's decision
    complicated the picture
    expecting the fed

24. punches klitschko briggs
    remembers opening his
    common one briggs

25. users stolen in
    today yahoo revealed
    affected users

4. Conclusion

Poems scored higher overall on their merits as text rather than summarization tools by a large margin as seen in table 4. Interestingly the kappa scores in table three indicates that we can be most confident about agreement between two ranker in terms of what they would deem “good summarizations” and “bad poems”. This is perhaps not so unexpected given the limited sample size and number ranking systems used. Variability occurs in higher numbers when participants were observing poems which are inherently linked at least somewhat to personal taste and preference unrelated to the task at hand. Simply put it’s hard to get two people to agree exactly what makes a “good” or a “neutral” poem but they know a bad one when they see it. Something we see again with the designation of “good summaries” which has the highest kappa score. It’s easy to tell when something has done a good job of capturing the essence of an article but further nuance is often rather subjective. Ultimately the haiku generator while effective was more useful as an amusement than a summarization tool.

References

Links used:

1. http://www.slate.com/articles/news_and_politics/politics/2016/12/john_m

ccain_and_lindsey_graham_are_not_your_friends_democrats.html

2. http://www.slate.com/articles/sports/sports_nut/2016/12/donald_trump_learned_his_political_moves_from_wwe.html?wpisrc=burger_bar
3. http://www.slate.com/articles/technology/future_tense/2016/12/how_russia_hacked_american_voters.html?wpisrc=burger_bar
4. http://www.slate.com/articles/news_and_politics/politics/2016/12/sleeping_giants_campaign_against_breitbart.html?wpisrc=burger_bar
5. http://www.slate.com/blogs/xx_factor/2016/12/13/pro_baseball_players_can_t_dress_like_women_in_dumb_hazing_ritual_anymore.html?wpisrc=burger_bar
6. http://www.nytimes.com/2016/12/13/us/politics/russia-hack-election-dnc.html?_r=0
7. http://www.esquire.com/news-politics/a51446/what-was-comey-thinking
8. http://www.npr.org/sections/thetwo-way/2016/12/14/505520902/planned-evacuation-of-east-aleppo-has-stalled-thousands-still-trapped
9. http://www.npr.org/sections/thetwo-way/2016/12/14/505512664/tech-leaders-to-meet-with-president-elect-trump
10. http://www.npr.org/sections/thetwo-way/2016/12/14/505482691/an-obama-backed-change-at-voice-of-america-has-trump-critics-worried
11. http://www.npr.org/sections/monkeysee/2016/12/14/505532188/if-tv-has-it-alan-thicke-probably-did-it
12. http://jezebel.com/woman-being-held-captive-escapes-after-writing-note-on-1790005467

13. http://jezebel.com/pipeline-dumps-176-000-gallons-of-crude-oil-into-a-nort-1790016454
14. http://www.npr.org/sections/thetwo-way/2016/12/12/505304688/lawsuit-challenges-fetal-burial-rule-in-texas
15. http://www.npr.org/2016/12/14/505581348/trumps-men-in-moscow-trump-disciples-suddenly-showing-up-in-russia
16. http://www.npr.org/2016/12/12/505286051/is-hillary-clinton-trying-to-question-the-legitimacy-of-donald-trump-winning
17. http://www.slate.com/articles/sports/sports_nut/2016/12/donald_trump_learned_his_political_moves_from_wwe.html?wpisrc=burger_bar
18. http://www.slate.com/articles/sports/sports/2016/12/the_english_soccer_sexual_abuse_scandal_is_like_penn_state_on_a_larger_scale.html?wpisrc=burger_bar
19. http://www.slate.com/blogs/xx_factor/2016/12/12/kesha_says_people_told_her_she_looked_better_while_she_battled_an_eating.html?wpisrc=burger_bar
20. http://themuse.jezebel.com/i-cant-decide-if-this-is-the-best-or-worst-christmas-al-1790017242
21. http://www.npr.org/sections/thesalt/2016/12/06/502172541/how-just-eight-flavors-have-defined-american-cuisine
22. http://www.npr.org/sections/thetwo-way/2016/12/14/505561759/final-arguments-set-to-begin-in-charleston-church-massacre-trial
23. http://www.npr.org/sections/thetwo-way/2016/12/14/505547771/interest-rate-hike-expected-wednesday-investors-await-move-by-federal-reserve

24. http://theundefeated.com/features/is-shannon-briggs-for-real/
25. http://gizmodo.com/yahoo-says-personal-info-of-over-one-billion-users-stol-1790119222