TextSpark

Introduction

- Submission for Apache Spark competition on DevPost
- Developed on DataBricks Community Edition

Business Use Case: Verbatim Coding

- Verbatim coding is the process of taking a set of user comments in freeform text and assigning them codes

- For example, a survey was conducted for Jupyter users

    - we would like to find out the common complaints, prioritize bug fixes and feature requests etc.

Business benefit

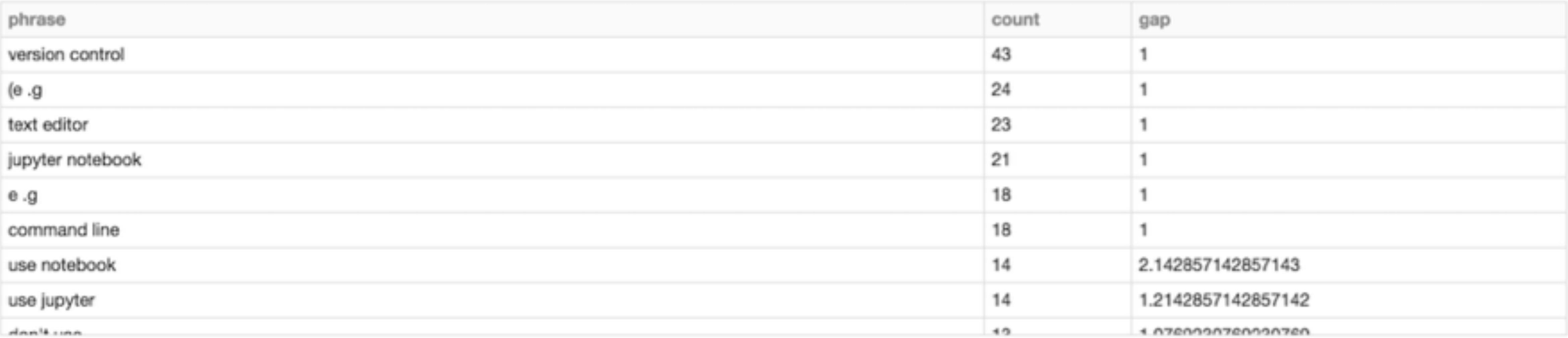
- Prioritize bug fixes and feature development
- In the case of general feedback from customers (e.g. surveys, customer emails) verbatim coding helps to
  - fix broken processes by addressing common complaints
  - provide direction for business development by identifying frequently requested features
  - can help identify potential at-risk customers by pattern matching against customer communications for customers who have already churned [1]

[1] Practical text analytics by Steven Struhl (2015)

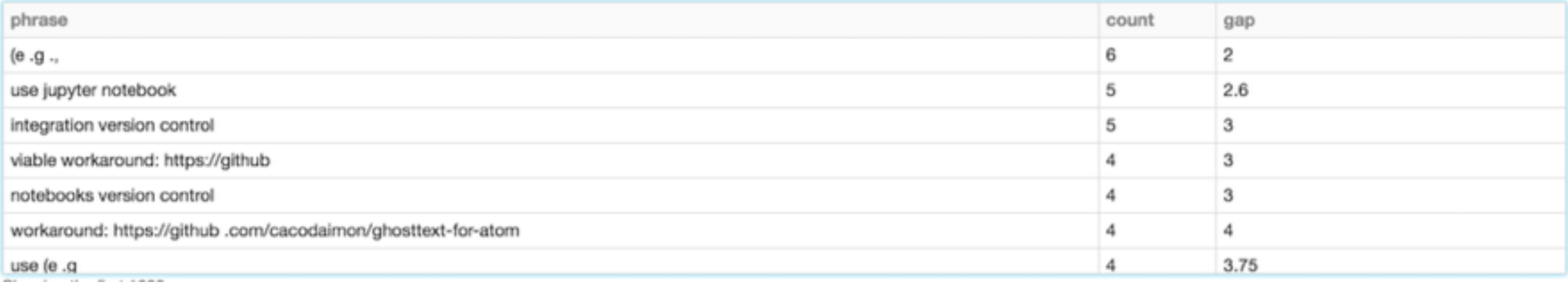
Survey data

- Jupyter survey from https://www.kaggle.com/jupyter/2015-notebook-ux-survey
- > 1700 rows of survey comments
- Some example freeform text fields in the survey:
    - What, if anything, hinders you from making Jupyter Notebook an even more regular part of your workflow?
    - Workflow Need #1:What needs in your workflow does Jupyter Notebook not address?

Example results



phrase	count	gap
version control	43	1
(e .g	24	1
text editor	23	1
jupyter notebook	21	1
e .g	18	1
command line	18	1
use notebook	14	2.142857142857143
use jupyter	14	1.2142857142857142
don't use	13	1.0769230769230769



phrase	count	gap
(e .g .,	6	2
use jupyter notebook	5	2.6
integration version control	5	3
viable workaround: https://github	4	3
notebooks version control	4	3
workaround: https://github .com/cacodaimon/ghosttext-for-atom	4	4
use (e .g	4	3.75

Showing the first 1000

Demo

- https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4033767068019271/830830361691472/95189476770593/latest.html

Future work

- Use ideas from NLP to improve the inverted index
  - sentence segmentation, stemming, entity recognition etc.