# Pink Team
# Interim Presentation

30/6/2015

James Doolan, Katharine Cooney, Kang Li,
Shuyu Huang, Liam Creagh

# Project Vision

- Personalise recommendation of news articles based on user's implicit preferences
- Generate implicit preferences from Twitter information

# Motivation

- Users are notoriously loth to express their preferences[1], in fact, it has been shown [2], that even when they do declare an interest in a set of topics these do not necessarily match their actual interests.

1.Doychev, Lawlor, et al. "An Analysis of Recommender Algorithms for Online News." CLEF, 2014.

2.Lavie, Talia, et al. "User attitudes towards news content personalization."*International journal of human-computer studies* 68.8 (2010): 483-495

# Motivation

- Instead of getting users to explicitly state their preferences, social media data is available that can be used to infer their interests. The accuracy of this method has been demonstrated [3].

3.*Bhattacharya, Parantapa/Muhammad Zafar/Niloy Ganguly/Saptarshi Ghosh/Krishna Gummadi* "Inferring user interests in the Twitter social network" (2014): 357–360. doi:10.1145/2645710.2645765

# Inspiration

[1] Bhattacharya, P., Zafar, M., Ganguly, N., Ghosh, S. and Gummadi, K. (2014). ***Inferring user interests in the Twitter social network***. p.357–360. [Online]. Available at: doi:10.1145/2645710.2645765

This paper describes the "crowd-sourcing" method of using Twitter user-defined lists for identifying "experts" on a topic. The inference is drawn that Twitter users following these "experts" are interested in the topic on which that person has been identified as "expert".

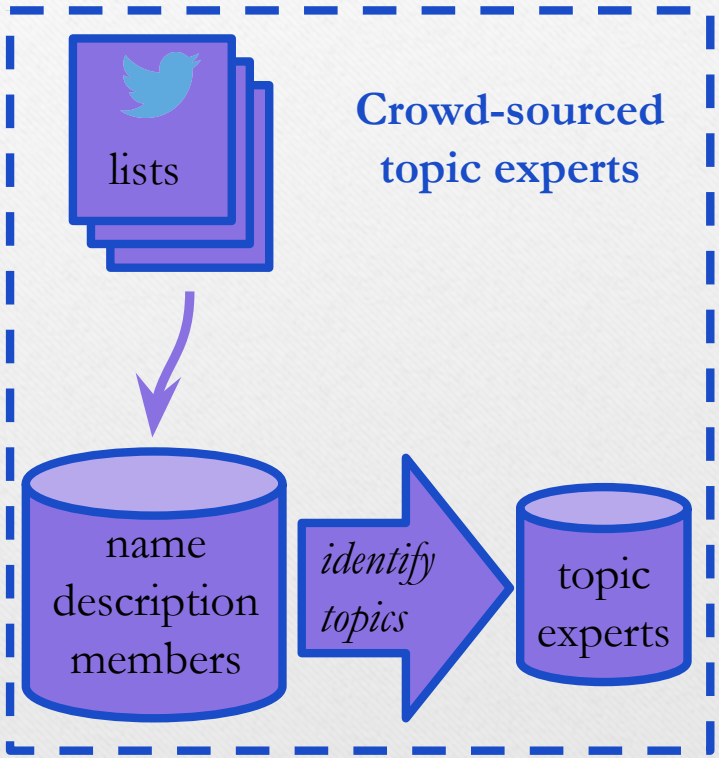
# Methodology
# Inferring user interests

- Twitter has a feature called “lists” which can be set up by any user
- Each list has a title and description and multiple members– these are provided by the creator
- A list is a set of accounts that share some common characteristic eg. comedians
- A public list can be followed by any Twitter user and accessed through the Twitter api

# Examples of Twitter Lists

| list_name | list_description |
|---|---|
| BeliebersIAdore | We are soo #WatitingForLollyVid gosh! Justin give us the fuckin' link! lolz |
| Football | |
| Real Madrid e Madridistas | |
| Football | |
| Futbolistas | |
| Popular Creative People | |
| Politics | |
| Theorising the City | |
| Media | |
| Innovasjonsfolk | |
| Social Good | |
| List 1 | |
| News | |
| Solutions Journalism | |
| Great accounts to follow | |
| Soljourno Newsrooms | Our partner newsrooms and other leaders in producing solutions journalism. |
| Global | |
| Philanthropy | |
| news | journalists and journalism |
| NA | North American voices |
| innovation | #innovation |
| the best | |
| Media Outlets / News | |
| whats-working | |

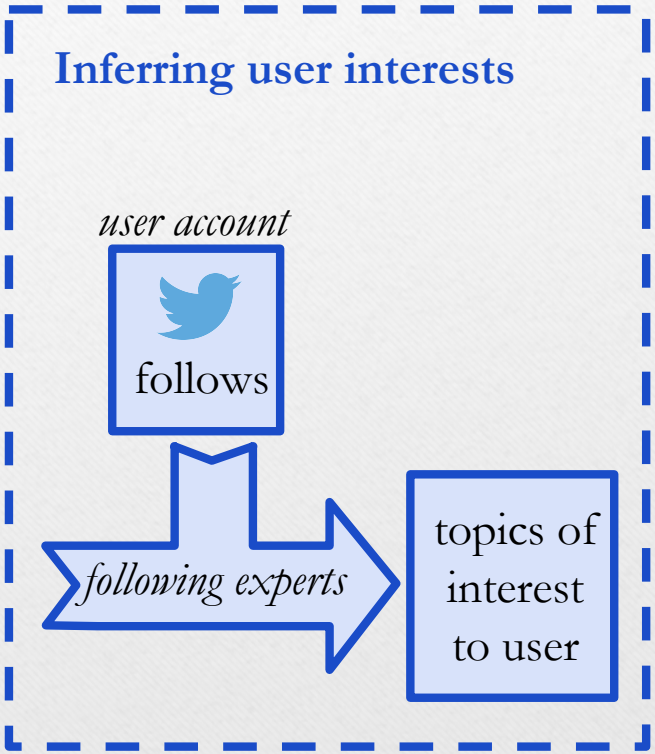
# Methodology (Final Product)
# Inferring topics and experts




*Topics* are identified using text processing of list names and list descriptions

*Topic experts* are identified as those with membership of >10 lists associated with a topic

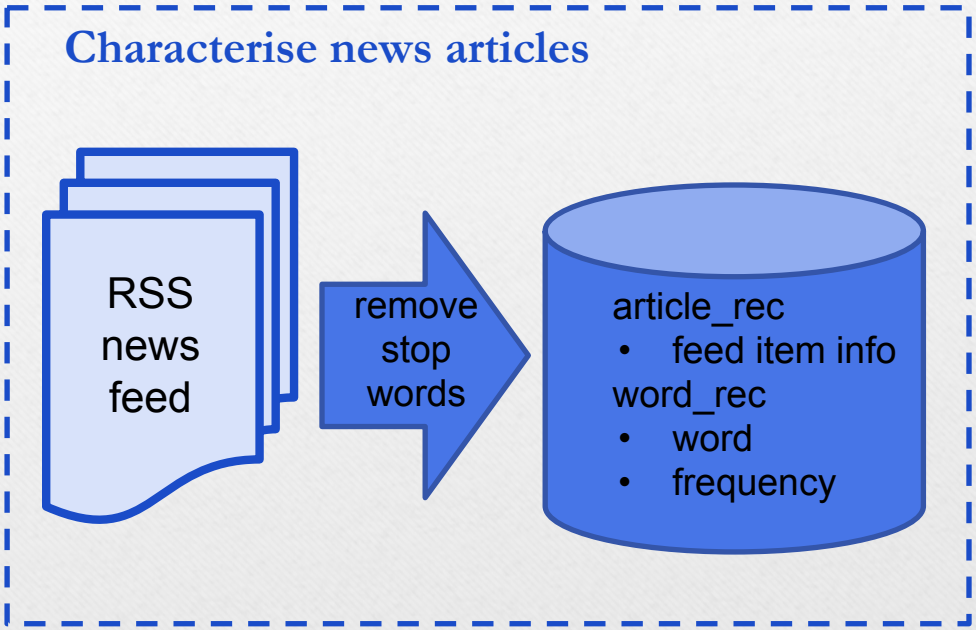
# Methodology (Final Product)
# Inferring user interests




Identify which accounts the user follows

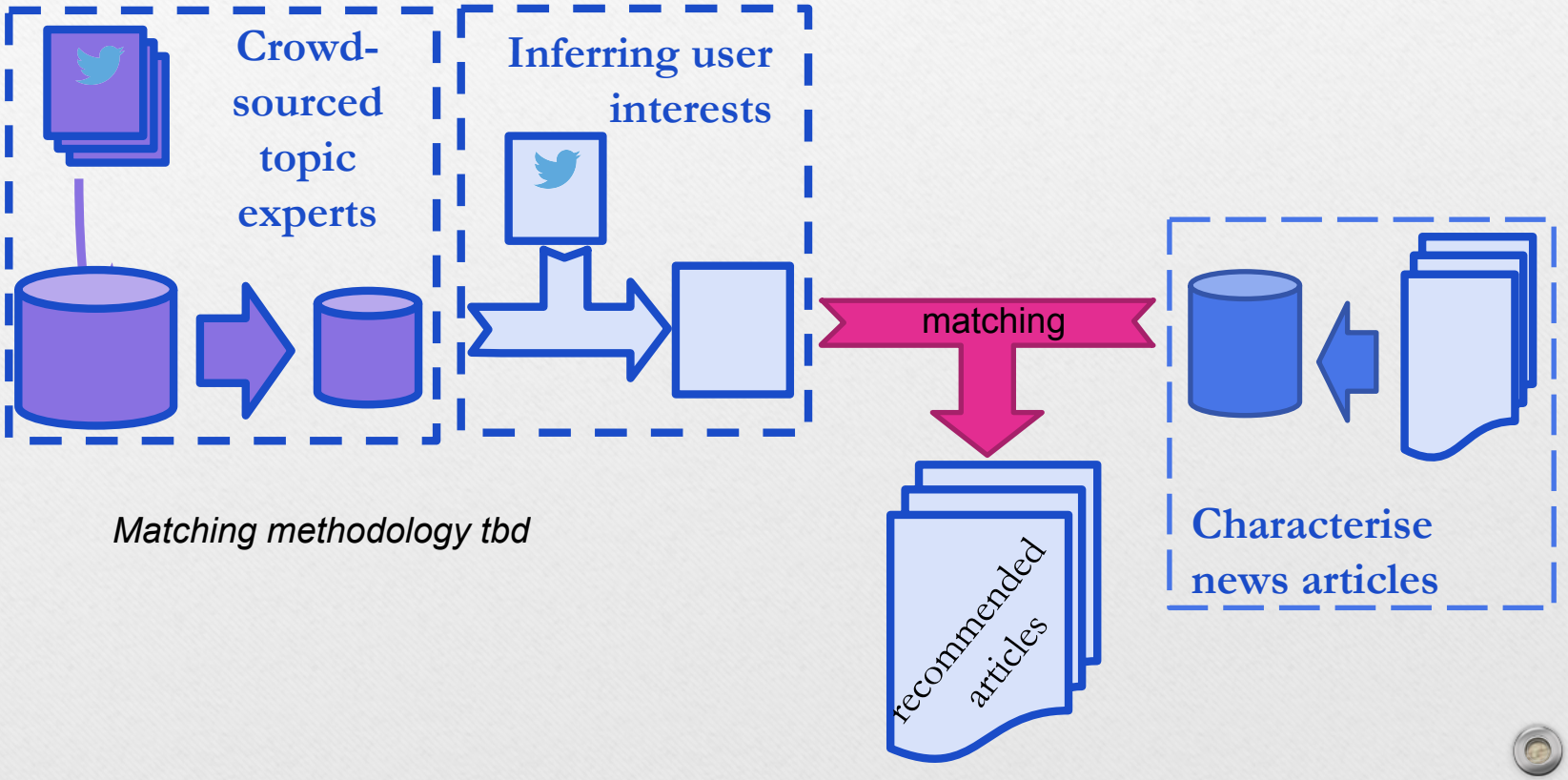
If this person is designated as an expert, infer that this user is interested in the expert's topics

# Methodology (Final Product)
# Characterising news articles




Word frequency records for each article in RSS feed are stored in database

# Methodology (Final Product)
# Personalised News Feed



# Possible pitfalls

- Identification of topics and "experts"
  - Many of the list names and descriptions used on Twitter are meaningless, as a result we need to parse large volumes of Twitter list to identify list topics and topic experts. The Twitter api is rate-limited, that is we can only extract a certain number of records every 15 minutes.
- User profile:
  - Although using implicit user data has been demonstrated to be more accurate than other methods of profiling users, this will only work where sufficient data is available for the user concerned. As we are using Twitter data to infer the user profile, this method will only work where a user is following sufficient number of people to make our inferences robust.

# Technology Stack

| Main Technologies | Languages | Python Libs | Other |
|---|---|---|---|
| Bootstrap | Python | Django Registration Redux | Android Studio |
| Django | HTML/CSS | Crispy Form | PyCharm |
| PostgreSQL | Javascript/Jquery | NLTK | Github |
| Android | SQL | Psycopg2 | |
| | Java / XML | Beautifulsoup | |
| | | Feedparser | |
| | | Tweepy | |

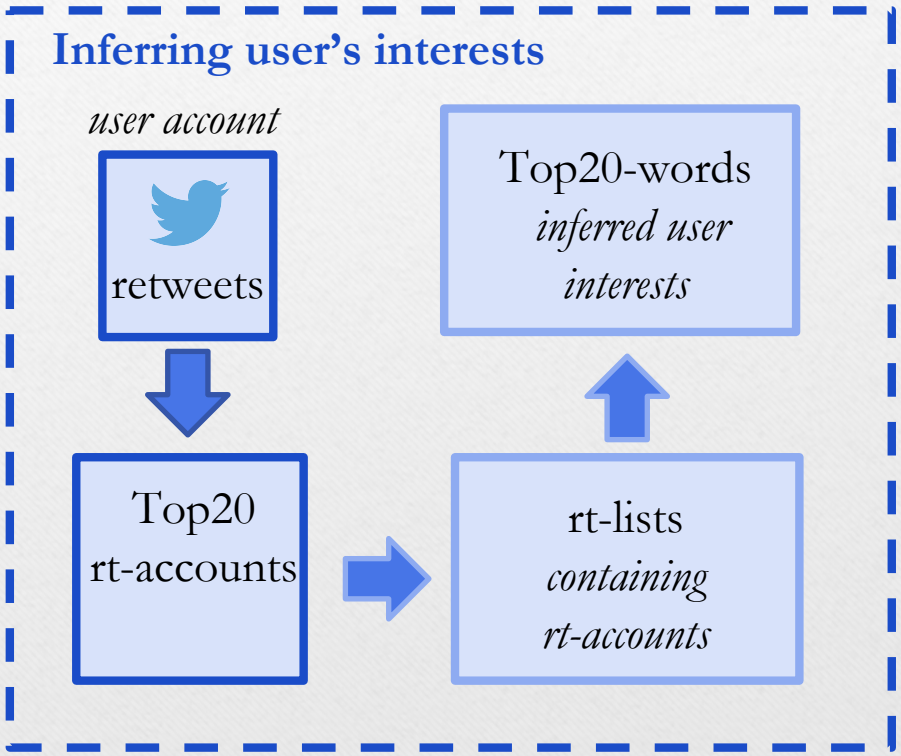
# Data Sources

- News sources:
  - RSS feeds from Reuters (initially)
  - Other RSS feeds (to be added) to add a broad range of views
- Twitter api:
  - Friendship ("following")
    - used to identify topics of interest
  - Lists
    - Used to identify topics and "experts"

# SWOT Analysis

| **Strengths** | **Weaknesses** |
| --- | --- |
| • potentially large userbase<br>• user appeal<br>• simplicity<br>• streamlined generation<br>• flexible<br>• modular<br>• extensible<br>• multiplicity of data sources available<br>• volume of Twitter data available<br>• # Twitter users | • large data – performance issues<br>• limited to active Twitter users<br>• dependence on Twitter lists<br>• dependence on Twitter<br>• difficulty of adding sources (as each is different) |
| Opportunities | Threats |
| • available via Android<br>• influence how news is consumed<br>• improve profile generation (additional social media)<br>• make recommendations more granular<br>• build larger news source<br>• create traffic for smaller news sites | • matching inaccuracy<br>• competition<br>• source dependent<br>• performance<br>• insufficient variety of sources to be appealing |

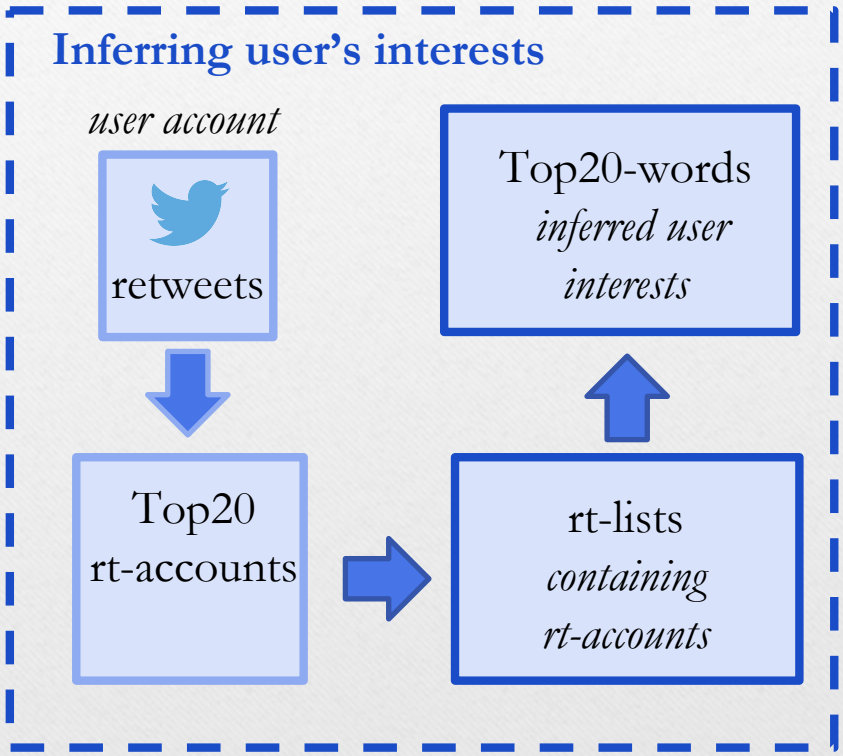
# MVP Methodology
# Inferring user interests(1)




User enters their Twitter screen name

The user's retweets are parsed to identify the top-20 accounts in their retweets ("top20 rt-accounts")

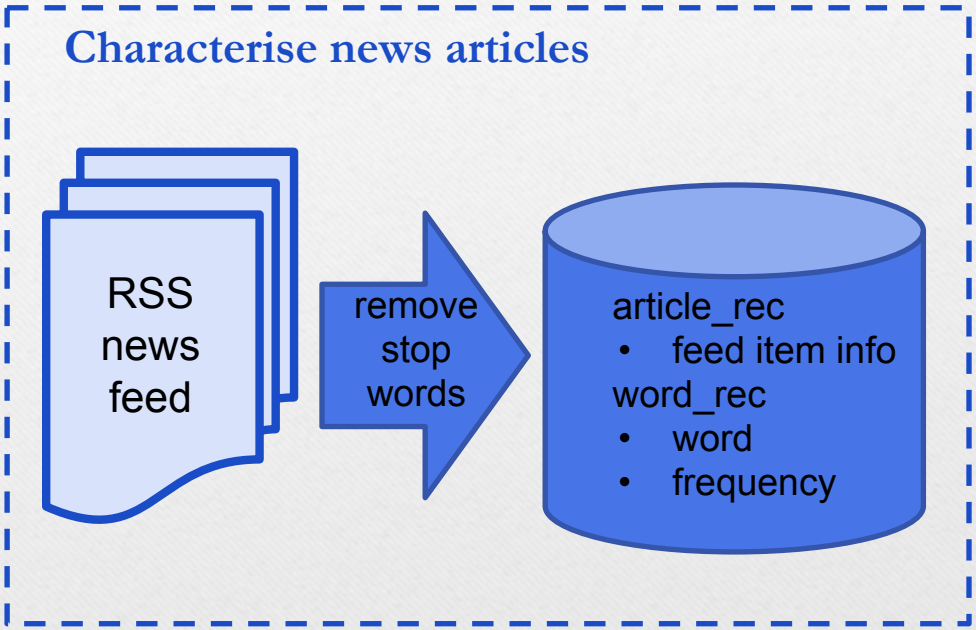
# MVP Methodology
# Inferring user interests(2)




For each of these rt-accounts we retrieve every list that that account is on – "rt-lists"

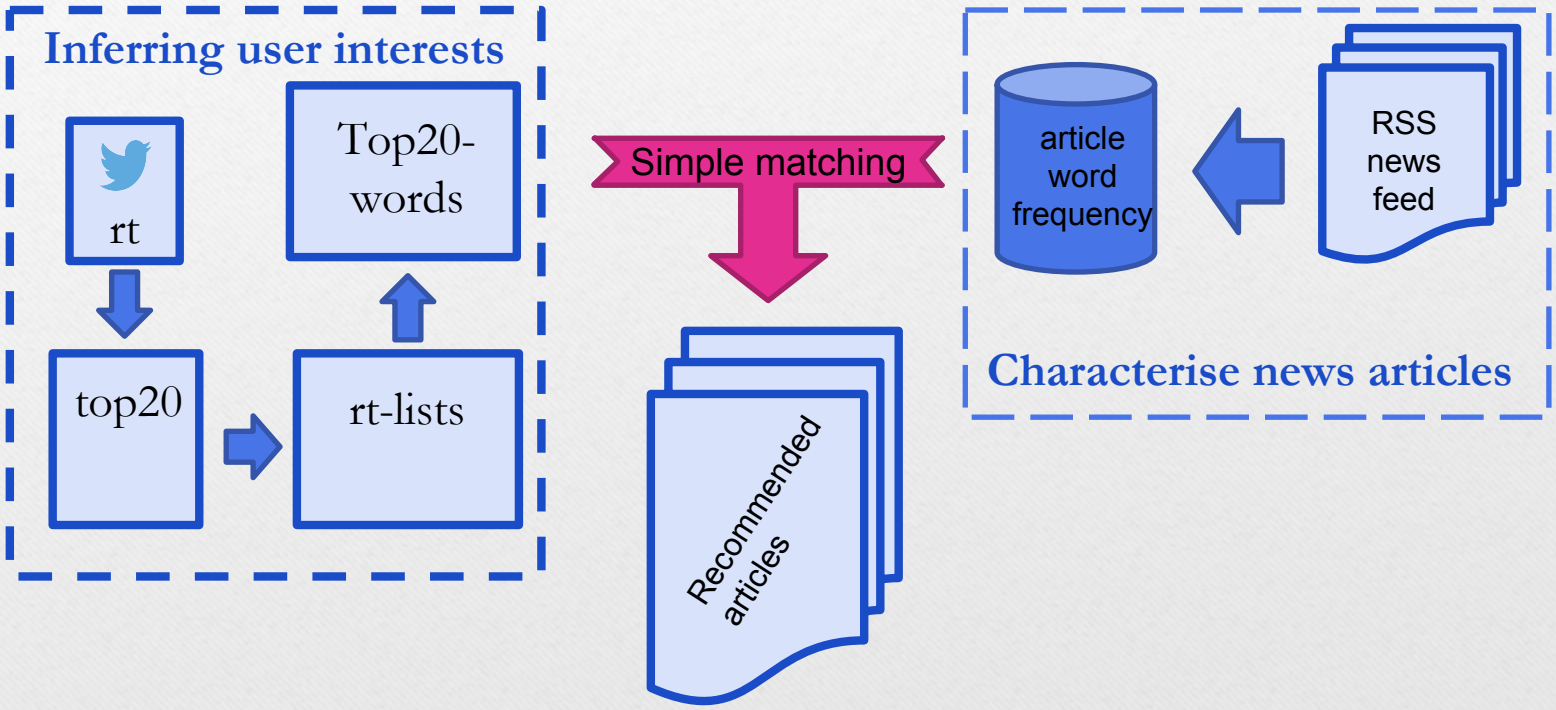
Retrieve the names of the "rt-lists" and generate a list of top-20 words associated with the Twitter screen name: "top20-words"

# MVP Methodology
# Characterising news articles




Word frequency records for each article in RSS feed are stored in database

# MVP Methodology



# Current Prototype Demo

- Our website will generate a feed of news articles aimed directly at you, based upon your particular tastes.
- You can do this by signing up to our site and then logging in to twitter to automatically generate your interests.
- Our news sources will be varied but will have diverse international content. Current: Reuters

# Immediate Next Steps & Project Roadmap

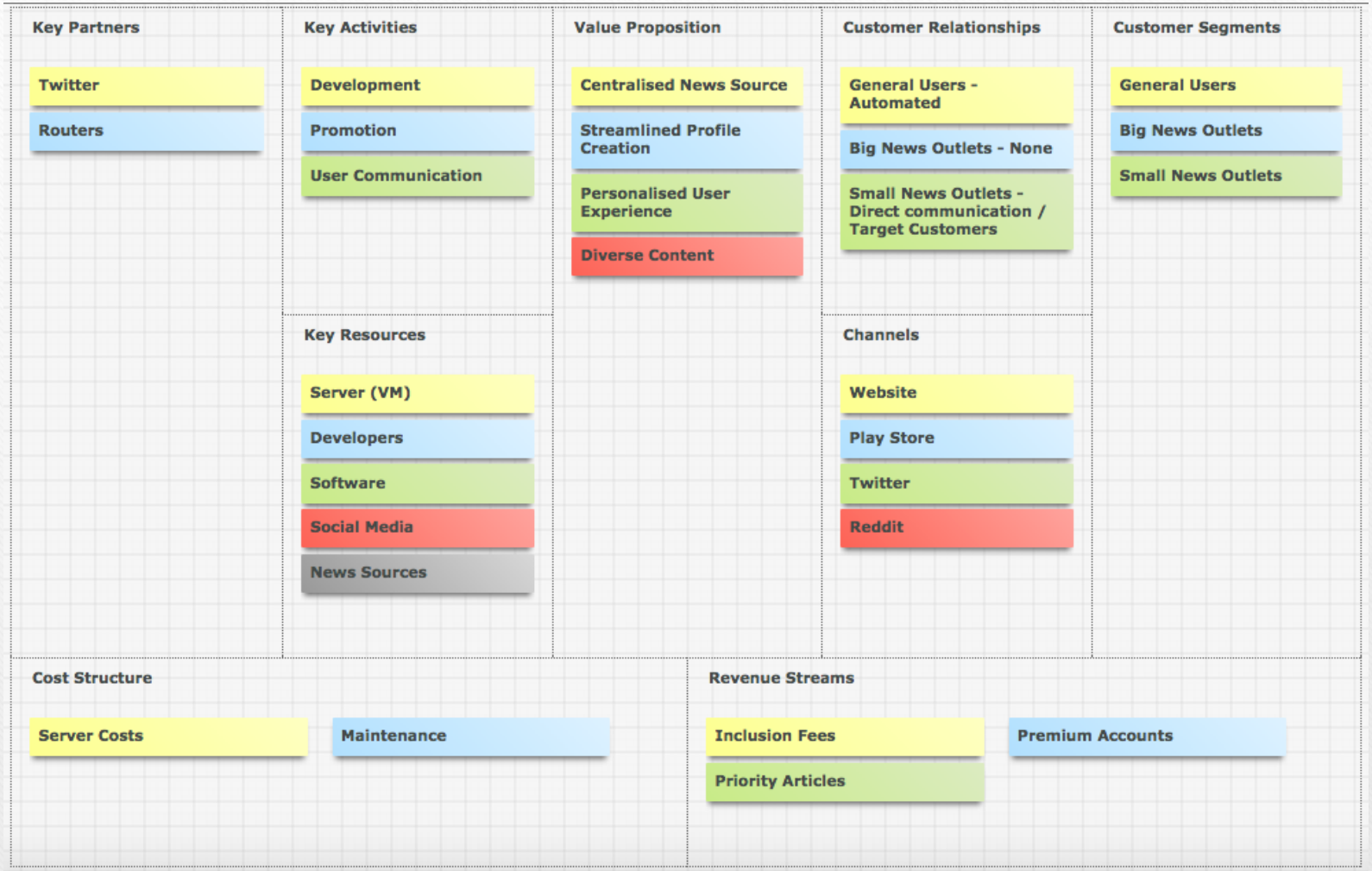
- Immediate next steps
  - Debrief from first 5 weeks' work
  - expanding data sources (Reddit, Pinterest)
  - Using goose extractor to standardise html parsing of articles – improve efficiency of adding sources
  - using list-experts methods to determine topics of interest
- Biggest challenge currently facing the project
  - Adding sufficient sources to make content relevant
  - performance – loading enough content to database for it to be accurate
  - Making real-time interactive performance acceptable to users

# Immediate Next Steps & Project Roadmap

- 2 week sprints:
  - Weeks 7-8
    - W7: UX survey (data gathering)
    - W7: create user stories/storyboards from quiz
    - Add Reddit as additional source for profiles
    - User testing
  - Weeks 9-10
    - W9: UX survey (feedback)
    - W9/10 add Pinterest
    - W10: replace html parsing (beautiful-soup) with goose-extractor
    - Make matching algorithm more sophisticated
    - User testing

# Immediate Next Steps & Project Roadmap

- Weeks 11-12
  - W12: benchmark survey with users
  - Display analytics report to user
  - Adding user functionality to explicitly change inferred interests
  - User testing
  - Evaluation




There is a two page accompanying document (Provisional Business Plan) explaining our Business Model Canvas in greater detail.

# User Experience

For UX we have created  2 surveys

- The first quiz will go out this week. (Week 6)
- This quiz is designed to help us create a variety of User Stories and fill in the following blanks:
- As a . . . . . I need . . . . So that . . . .
- It first describes our proposed prototype and asks the user various questions about themselves and their thoughts on the prototype.

  Available here: https://www.surveymonkey.com/s/FM2WMYB

# User Experience

- The second quiz will go out in week 9 and again in week 12
- This quiz is designed to help us find out what aspects the site can be improved on and how the user feels about the different aspect of the site
- It will also help us benchmark our progress by averaging the results between the 2 different times we send out the quiz
- There are 10 questions and each question is answered on a scale of 1 to 10.
- The questions will be based on Nielsen's usability heuristics