Pipeline plan

Alex Bishop

October 7, 2020

Contents

1 Pipeline planning

1.1 Policy questions

1. What do business website notices say about exposure/response?
   (a) Topic evolution over: time, space, industry
   (b) Where are there more or less notices than expected?
   (c) Sentiment analysis [HARD]

2. What do social media posting say about exposure/response?
   (a) How often are social media posts updated?
   (b) Longevity - when do accounts start and stop posting?

   (c) Sentiment analysis [HARD]

3. What are the levels of adoption of technologies/strategies such as e-commerce?

   (a) Are there changes in the glass e-commerce flag [DQ ISSUES]

   (b) Detect a change of strategy using NLP on notices or social media activity[HARD]

4. Sectoral and geographic distribution of exposure

   (a) Based on trends

5. What are the diversification opportunities

   (a) Sector co-occurrence

6. What is the exposure to Covid of innovative companies

   (a) Compare the effect of prescence in datasets

7. Relationship between exposure and business failure?

8. Can we identify businesses at risk of failure?

   (a) Industry + place exposed

   (b) Innovative flag

   (c) e-commerce / new strategy or technology (Point 3)

   (d) Covid notice containing "negative" information/outcome

9. Indirect impacts - OUT OF SCOPE

10. Can we detect and measure business failures in a timely manner?

    (a) Explore the lag and accuracy of companies house [DATA ACCESS ISSUES]

    (b) Explore alternate datasets such as:

       - `https://www.gov.uk/government/statistics/incorporated-companies-in-the-uk-`
       - `https://www.gov.uk/government/statistics/monthly-insolvency-statistics-jun`

1.2 Data issues

No steps are explicitly dependent on these points, but time invested into these will correspond to an increase in data quality.

11. How to deal with multi-site firms
    (a) How to resolve conflicting information between trading address (Companies House) and the addresses (up to 5) found by Glass AI.
    (b) How to discern a true multi-site firm from a false one (e.g. a second address is detected that corresponds to e.g. the address where an event is to be held)
    (c) What if there are more than 5 sites?
12. What is the coverage and bias of the business website data?
13. What is the most reliable way to match industrial taxonomies? Glass AI uses Linkedin's industry list, Companies House uses SIC codes, and we wish to report SIC codes.

1.3 Engineering "tasks"

14. Data modelling of Glass datasets, including the consistent merging of multiple snapshots
15. Fuzzy-matching
    (a) Develop reusable pipeline
    (b) How to validate accuracy of fuzzy-matching?
16. Finding social media links - requires Nesta to scrape websites of organisations. NOTE: legal ramifications of scraping facebook data limit us to using Twitter only.
    (a) Do two small pilots:
        i. Scrape and retrieve twitter handles from business websites in location X - e.g. one local authority
        ii. Scrape and retrieve twitter handles from business websites in industry Y - e.g. one SIC code
17. Fetch twitter data for organisations with social media links.

(a) Explore Twitter API costs and rate-limits

   - Every 24 hours we can fetch ~100,000 tweets which is sufficient for a pilot stage but would be insufficient to collect historic tweets (there are likely >100,000 new tweets being generated a day across the organisations within Glass)
     - Limitation: you can retrieve the last 3,200 tweets from a user timeline, therefore if they tweet frequently then we may not be able to far back enough in time.
   - There is the potential to apply for an academic research account which would give better access for free: `https://developer.twitter.com/en/solutions/academic-research`
   - The "Enterprise" pricing of paid access to the Twitter API (which allows full historical access rather than the last 30 days or last 3,200 tweets) is unclear but costs at least $99/month

(b) Building on previous point:

   i. Fetch and store data from Twitter API

   ii. Generate simple update statistics for each profile - i.e. ignore textual content - and assess utility of data initially based on coverage and frequency

   iii. Analyse textual content of twitter data [SCALE-UP]

1.4 Dependencies

1. [14]
2. [16, 17]
3. [14, 1]
4. [15,]
5. [15,]
6. [15, 1, 2, 4, 10] (Relies on some measure of exposure)
7. [1, 2, 4, 10, 3] (Relies on some measure of exposure)
8. [5, 6, 7]
9. OUT OF SCOPE
10. None

11. [14, 15]

12. [14]

13. [14, 15]

14. None

15. None

16. [14]

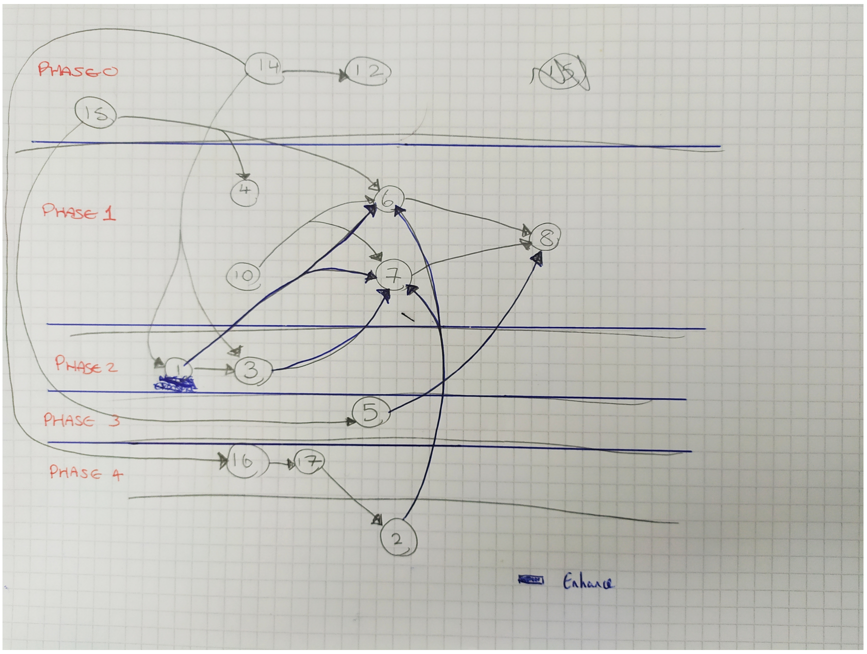
17. [16]

1.5 Task graph

1.6 Proposed task order

Point 8 is the end-goal in the sense that it combines all the previous components of the analysis. Achieving a first iteration of 8 could be seen as our MVP.

Not all of the components are required to yield a MVP of 8.



Note: 6 could be skipped if not deemed important by SG - it adds significant complexity by introducing datasets not used elsewhere (PATSTAT and Gateway to Research)

1.6.1 Phase 0 - Groundwork

Exploratory analysis: [14, 15, 12]

1.6.2 Phase 1 - MVP

Fastest path to 8: [4, 6, 10, 7]

1.6.3 Phase 2 - Iteration

Enhance 8 with 1 and 3 (via. 6 and 7)

1.6.4 Phase 3 - Iteration

Enhance 8 with 5 (Diversification opportunities)

1.6.5 Phase 4 - Social media pilot

16 and 17 (allow 2 to be performed)

1.6.6 Phase 5 - Iteration

Enhance 8 with 2 (via. 6 and 7)