{ Codebook for Web Classification }

Please note that the scheme of classification specified in this Codebook is more granular and expansive than the scheme of Base Format outlined in the manuscript. In accordance with our conceptual framework, we combined a couple of categories as post hoc adjustment.

This codebook is for coding an individual website and submitting relevant information via a google form consisting of two pages.

If you want to revise your coding of the website just coded, click "edit your response" immediately upon submission of the form. However, if you realize later that you need to modify codes for some websites, submit a new form again for that website, using the same website ID Number and Website Domain. For each website, we will consider only your latest submission.

Sometimes a website is temporarily down. Thus after you finish coding all the websites, please try again to access the ones that were inactive.

Lastly, as this is an ongoing research project, all information/data involved should be kept strictly confidential.

Coder Name Choose your name from the dropdown menu.

Website ID Number Enter the number corresponding to the spreadsheet (eg. 536).

Website Domain Copy from the spreadsheet (eg. douban.com)

Base Category

Choose one from the following 9 base categories:

```
1 Search Engine
```

A search engine allows users to identify and navigate things hosted outside of the search engine itself (eg. Google, Baidu).

```
2 Social Network
```

Social networks are where users are able to access content based on their network (friends, connections, who/what they follow) on the sites (eg. Facebook, Pinterest, LinkedIn, Twitter).

- Blogging Platforms without a "portal" homepage (see below for detailed definition) such as Wordpress.com are Social Networks.

Portal curates large amounts of symbolic content from diverse sources that are not produced in a centralized way by the site itself. Examples of portals include :

- General portals such as Yahoo and MSN (and more country specific ones such as Naver, Mail.Ru, and Sina).
- Review Websites such as Yelp and IMDB; Wikis including wikipedia; Blogging Platforms such as Livejournal.com, which curate diverse blog content at its homeage. These sites display and organize content generated by numerous users.

- More specialized portals such as job portals (eg. Monster.com and 51job.com) and education portals (eg. coursera.org; however, note that single university websites providing the university's courses should not be considered as Portals).

```
3 Portal - Having news
```

Some portals include news, even though news may not be their "primary focus." Having news is defined as featuring timely information/commentaries about current affairs in the public domain catering to a general audience. Examples of "Portal - Having news" include:

- MSN.com, Yahoo, Sfglobe.com and Youtube.

```
4 Portal - Not having news
```

These are portals that provide information that is either in private/leisurely/consumerist/highly specialized domains, or not time sensitive, (or both).Examples of "Portal - Not having news" include:

- Webmd, mayoclinic, allRecipe, coursera, wikipedia.org, tripadvisor, yelp.

Singular Content Producers / Publishers are like Portals in managing symbolic content (in contrast to material goods). The key difference from Portals is that they have their own content (generally produced through own staff of pre-contracted content contributors such as columnists and studios). Some examples to illustrate this category:

- ABC.com, the online extension of TV channels, goes here, but Youtube and Vimeo would be Portals
- Techcrunch would fall in this category, although you may be tempted to call it a Portal (A careful look reveals that they have a fixed staff and a regular fixed set of contributors who create all the content).

```
5 Singular Content Producer / Publisher - Having news
```

These Singular Content Producers / Publishers feature timely information/commentaries about current affairs in the public domain catering to a general audience. Examples include:

- Wired magazine, The New Yorker, Huffington Post, and Forbes.

```
6 Singular Content Producer / Publisher - Just information / broadcasting entertainment
```

These Singular Content Producers / Publishers carry content that is either in private/leisurely/consumerist domains, or not time sensitive, or both. Examples include:

- Techcrunch, CNET, and Harvard Business Review (vs Forbes).

E-Commerce / Retail sites deal with "material" or "utilitarian goods" rather than "symbolic" or "experiential goods" (the latter are what "Portals" and "Singular Content Producers / Publishers" deal with). E-Commerce / Retail directly charge ordinary users in exchange of their products, although sometimes they have free and premium accounts (eg. Dropbox). Some examples that are NOT E-Commerce / Retail to help better understand this category:

- Sites that distribute/sell videos/movies/games such as Netflix, Hulu and Steam are Portals (symbolic content)

- Online education sites (even if they charge money) deal with symbolic goods and thus either Portals (eg. Coursera) or Singular Content Producers / Publishers (HBX by Harvard Business School).
- For the same reason, information brokers/enablers such as Job portals, Craigslists, etc. are also to be classified as Portals rather than E-Commerce sites.

```
7 E-Commerce / Retail: Centralized distributor of diverse products produced by others
```

Sites themselves (primarily) are centralized distributors of the diverse products initially produced by others (e.g. Amazon and Barnes & Noble). In other words, the website's brand (eg. Amazon or Barnes & Noble) is not associated with the products they sell.

```
8 E-Commerce / Retail: Distributing its own brands
```

These sites produce and sell their own "brands" or "services" (eg. Apple, Nike, Samsung, dropbox, 365rili [Chinese online calendar], Paypal).

```
9 E-Commerce / Retail: platforms hosting sellers other than themselves
```

Websites under this category are online platforms hosting numerous sellers or service providers (eg. ebay, airbnb, Uber). The platform facilitates transactions; the individual sellers operating on the website, rather than the website itself, have the rights to close deals.

Other Attributes (Relevant for all websites)

Check one of the two boxes for the following two items:

```
1 Driven by user-generated content (UGC)
0 Non-UGC driven
```

Following features of the site make it UGC:

1. Creation outside of professional routines and practices – w/o the expectation of profit or remuneration (eg. "online literature" sites selling works by "amateur writers" to readers, or websites selling photos by "amateur photographers" are not UGC-driven).
2. Creative effort. Therefore a news website wouldn't be coded as UGC site just because it allows users' comments.
3. (Semi-) public distribution (eg. Facebook is but Skype and WeChat are not).

Other examples for illustration:

- Wordpress.com (the blogging platform) is a UGC site. But Wordpress.org (the site which offers tools for making websites to be hosted anywhere) is not a UGC site, as the latter does not have UGC content on its domain.
- In general, websites that offer "services" to people to host/build websites etc are not UGC sites (e.g. 110mb.com) because (a) these sites are indifferent to and independent of the actual content that their users produce, and (b) their users' sites (built and hosted on these sites) are not necessarily UGC either.

```
1 Legacy: online extension of existing offline operations
0 Purely online
```

Legacy sites are the online extensions of existing offline operations. Examples include:

- Many websites by traditional media, i.e., traditional content providers / publishers/ broadcasters who have news and/or entertainment content.

Purely Online is where no ownership/direct affiliation can be mapped to an offline entity. In certain cases, ownership may be tied to an existing offline company, but the online product may not have a clear predecessor. For instance

- MSN belongs to Microsoft, but it is clearly a new online product. Therefore MSN.com is "online only", but microsoft.com is legacy as that site is an online presence of the Microrsoft Corporation.

Name of Website (optional) Popular names of websites which may not be evident with the domain name (more relevant for non-English/foreign websites).

Remark (optional but recommended) Note important keywords that pertain to a site's most characteristic feature/association, separated by ";". These keywords may be related to the themes that the website is focused on, and/or to certain specific functions that cannot be captured by our existing standardized categories. For example,

- For IMDb.com, remarks may include: movies; reviews