POD**CATS:** A/B Testing

ONLINE TEST PREFACE

Figuring out how to do our A/B testing as accurately as possible was an extremely challenging task. This is because legitimate use of our application required an actual live lecture. Otherwise, user actions on the application would really just be dummy actions and would not reflect actual use. Our solution to overcome this obstacle is detailed below in 'online test setup.'

ONLINE TEST SETUP

Our solution was to take a segment from a real lecture (podcast.ucsd.edu) and play that for a user and have them use the application in the simulated lecture environment. The segment used was 0:00-5:00 of CHEM6A - LE [B00] - Professor Hoeger - Winter 2015 - Monday 2/9/15 Lecture. During this portion of the lecture, the professor started the lecture, covered exam info, logistics, and an example problem. This was selectively chosen because of the variety of information presented within a short 5 minute period. All users being tested watched this segment of a lecture and were instructed to use the application as if they were actually in that lecture.

A/B VERSION SETUP

As for the A/B setup, the two different versions were:

1. Bookmarks without tags (original), version A
2. Bookmarks with tags and comments suggested as optional (experimental), version B

The version determination was selected by a *random number generator* with a 50/50 chance of landing on either version. We received a pretty decent spread with 9 users trying out version A and 11 users trying out version B.

ONLINE TEST RESULTS

https://docs.google.com/spreadsheets/d/1ugdfKaQv-9BAuUpreYN-2LS7Nbyst0fYvHjH8fh-XoE/edit?usp=sharing

What we hoped to determine from these test results is whether or not one version is better than the other at guiding the user into knowing what the structure of the application (PodCats) is and how it should be used.

QUANTITATIVE ANALYSIS

Preface: Chi-Squared is not applicable to use in our quantitative analysis as we measured number of actions per user, rather than if a user performed a single action or not based on the version. For our purposes, we found that a better measure, in line with the goal of the experiment, would be to use the *standard deviation* for number of bookmarks to see how widely the use of bookmarks varied between users in each version group.

Of Bookmarks:

1. Version A: 15, 6, 7, 9, 6, 6, 8, 7, 7,
    a. Standard Deviation: 2.848
2. Version B: 4, 3, 3, 4, 4, 3, 5, 3, 3, 4, 5
    a. Standard Deviation: 0.786:

Looks like version B had a significantly smaller standard deviation than version A, indicating that the number of bookmarks that version B users used varied a lot less across that group than in version A. The interpretation of this quantitative analysis is that version B better guided users into how they should be utilizing the application, where as version A left the use of the application a lot more open ended, thus resulting in a larger variety in bookmark numbers across users in that group.

Of Edits:

1. Version A: 2, 1, 1, 1, 1, 2, 1, 0, 2
2. Version B: 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1

We determined that the number of edits was insignificant data in our experiment because in the time frame of 5 minutes, most users were simply hitting the edit button once to test it out, rather than actually use it for any editing purposes. Thus, this data point is not useful in determining the goal as mentioned above under section 'online test results.'

Of Deletes:

1. Version A: 0, 1, 2, 1, 1, 1, 1, 0, 3
2. Version B: 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1

We determined that the number of deletes was insignificant data in our experiment because in the time frame of 5 minutes, most users were simply hitting the delete button once to test it out, rather than actually use it for any editing purposes. Thus, this data point is not useful in determining the goal as mentioned above under section 'online test results.'

QUALITATIVE ANALYSIS

Observed data: Content of bookmarks.

Version A:

Across all the users testing version A, the bookmarks are pretty descriptive and generally share some of the key highlights in that lecture. Namely, that the exams were ready for pickup in York 4030 and that the lecture was starting on the topic of polyelectronic atoms. One of the key observations with the version A group is that based on the bookmark content, the variation of use is pretty large. Some users left detailed comments, some left less detailed comments. The one commonality though is that the bookmarks seemed to be used notes themselves, almost like the content someone would put in a notebook.

Version B:

Across the users testing version B, the bookmarks are actually very similar. Users in the version B group all have bookmarks that follow this same tagged structure: Lecture Start, Exam Info, Logistics, Example Problems (with the exception that some users omitted the Logistics tag). This shows qualitatively that version B with tags seems to offer users a more structured guide in how they should be using the application. Another important note is that in version B, the comments (if any provided), are more like titles for a section in the lecture, rather than full blown notes like in version A. This will be an important distinction between version A and version B to be analyzed in the summary of key findings.

SUMMARY OF KEY FINDINGS

The overall result appears to indicate that version B, the version with tags, does indeed better guide the user into knowing what the structure of the application is and how it should be used.

The quantitative analysis showed that version A users had a much larger standard deviation in number of bookmarks issued, showing that the intuition between each user in the group had a much higher chance of variation. This correlates to a less guiding structure for the version A application page. On the flip, version B users had a much smaller standard deviation in number of bookmarks issued, showing that the intuition between each user in the group had a less chance of variation. This in turn correlates to a more guiding structure as users seemed to all understand and act in similar ways.

The qualitative analysis shows that version A users were more so using the application as just another form of taking notes. The comments of the version A bookmarks closely resembled text that you would continuously jot down on a notebook; this is not what we were aiming for in the design and use of the application. Version B users, on the other hand, had bookmarks with comments that were more used as titles of the tagged interval. This more so actually represented the 'bookmark' concept we had envisioned from the get go. Therefore, the qualitative analysis seemed to indicate that version B was better in guiding the user toward the overarching function of our application PodCats.

In conclusion of the key findings, the data all appeared to point at version B over version A as being the better page to use for our live lecture bookmarking page.

LIST OF POTENTIAL REVISIONS

- Change design of podcatsit page from Version A (current) to Version B
    - With that, known Version B UI bugs need to be fixed (see Dev Plan)
- For Version B, consider changing "optional comment" to other wording to make it even more clear what the comment is supposed to be used for if we want to nudge users towards a specific use (and away from the notebook use case)
- Allow users to customize or add to the set of predefined tags in the settings page to give more flexibility while still offering the nudged structure benefit of version B. This is sort of a combination of the positives of open-endedness and the positives of nudging.
- Reconsider the character limit (number to be determined) in a best effort to again nudge the user into how they should be utilizing the comment box. It seemed from the data and user testing that 100 characters was still pretty long. As we could tell from the version A online testing results, this gave enough freedom for many users to interpret that box to simply be used as if they were taking notes in a notebook. (100 characters may have not been restrictive enough to warrant any different reaction/ intuition).
- Overall layout of bookmarks (UI/UX) → Uncoupling the submission of tags and optional comments to prevent the users from making unwanted ("No Comment") comments
- Switch tags into their own separate data component of a bookmark, rather than group with comments (basically un-wizard that part of version B)
- Keep checking for mobile incompatibilities and update accordingly