

Checkpoint V: Report

Diogo Nicolau
IST
Lisboa, Portugal
digasnikas@gmail.com

Filipe Teixeira
IST
Lisboa, Portugal
filipelteixeira@gmail.com

João Moura
IST
Lisboa, Portugal
joacostamoura@tecnico.ulisboa.pt

ABSTRACT

Nas seguintes páginas encontra-se em detalhe um report do trabalho desenvolvido pelo grupo 27, na cadeira de Visualização de Informação - Alameda. Para o efeito serão abordados vários tópicos lecionados na cadeira, essenciais para a conclusão do projeto. O report está organizado segundo o planeamento da cadeira, mantendo assim o registo temporal de todas as aulas teóricas e práticas.

O objetivo é dar ao leitor conhecimento de todos os passos que o grupo realizou para a implementação do projeto, que consiste na criação de uma visualização de um conjunto de dados recorrendo a tecnologia web e a bibliotecas como a d3.js.

Keywords

Visualização de Informação; Stack Overflow; Stack Exchange; Análise de Dados; D3js; DataSets; Técnicas de Visualização; Interação.

INTRODUÇÃO

A escolha do domínio foi feita baseada no tamanho do dataset e na estrutura já existente para a recolha de dados. O Stack Overflow é um site/plataforma de dúvidas técnicas sobre informática. Os utilizadores podem fazer vários posts, desde perguntas, respostas ou simples publicações. O grande número de utilizadores ativos torna a comunidade de Stack Overflow um caso de estudo, sendo a chave do sucesso desta plataforma. Naturalmente esta imensa atividade gera um enorme flow de informação. Acontece que toda esta informação fica guardada em bases de dados. Motivados pela necessidade de partilhar com os seus utilizadores o maior volume de dados possíveis, a equipa do Stack Overflow criou um *data explorer* no Stack Exchange. Este *data explorer* engloba dados de várias plataformas como *Mathematics*, *Wordpress* ou *Salesforce*. Para aceder aos dados escrevem-se várias queries em T-SQL. Uma vez composta a query basta correr-la e, caso esta não apresente erros de sintaxe ou de lógica, o *data explorer* devolve um conjunto de dados que podem ser exportados na extensão .csv. A estrutura da base de dados desta informação é complexa, dado que modela por inteiro todo o domínio do Stack Overflow. Sempre que um utilizador escreve um post ficam registados vários traços desse mesmo utilizador, desde a data e hora do post ao local onde o post foi escrito. O interesse neste domínio está na possibilidade de existirem padrões não perceptíveis na comunidade Stack Overflow, e com a informação existente

extrair e perceber as características de informáticos em todo o globo. Dito isto o grupo pretende, através da VI, realizar diferentes tarefas que facilitem e revelem informações como:

- Países com maior número de perguntas e respostas, e correlações entre estas;
- Tecnologias mais utilizadas;
- Características específicas que diferem entre os informáticos de cada continente;
- Crescimento na utilização de diferentes tecnologias;

Dito isto o grupo comprometeu-se desde cedo a realizar as seguintes tarefas:

- ❖ **Task 1:** explorar a correlação entre o número de utilizadores e o número de perguntas, respostas e top utilizadores dum país;
- ❖ **Task 2:** apresentar as tags mais utilizadas e o seu crescimento ao longo do tempo;
- ❖ **Task 3:** sumarizar as perguntas e respostas de um utilizador ao longo do tempo;
- ❖ **Task 4:** procurar os países com maior número de top users, e identificar as tags utilizadas por estes users;

Apesar de bem definidas as tarefas o grupo sentiu mais tarde a necessidade de alterar em parte algumas das tarefas e inclusive remover outras. Considerámos que entrar no domínio de utilizadores específicos iria atrasar o projeto, e optámos por não abordar as tasks 3 e 4. Incluímos novas tasks que se enquadram melhor com as já existentes, focando agora mais no registo espacial. No final ficámos com as seguintes tarefas:

- ❖ **Task 1:** explorar a correlação entre o número de utilizadores e o número de perguntas, respostas e top utilizadores dum país;
- ❖ **Task 2:** apresentar as tags mais utilizadas e o seu crescimento ao longo do tempo;
- ❖ **Task 3:** sumarizar a utilização de tags por continente;
- ❖ **Task 4:** apresentar os países com maior número de perguntas, respostas e top users;

Estas tasks serão dissecadas mais tarde ao longo deste relatório, onde pretendemos argumentar e justificar algumas das decisões tomadas pelo grupo em volta destas tarefas.

Copyrights are held by the authors:

- Diogo Nicolau;
- Filipe Teixeira;
- João Moura;

TRABALHOS RELACIONADOS

Existem sem sombra de dúvida muitas VI's e infogramas de teor informático. São realizados diariamente estudos sobre várias tecnologias de forma a entender as novas tendências. A indústria IT é provavelmente a indústria que mais rapidamente se transfigura, e naturalmente estas empresas investem grandes quantias de dinheiro de forma a saberem qual será a próxima grande tecnologia e assim estar em vantagem competitiva. Encontramos online várias VI's para o domínio das linguagens de programação:

A figura 1 em baixo apresenta as linguagens de programação mais predominantes;

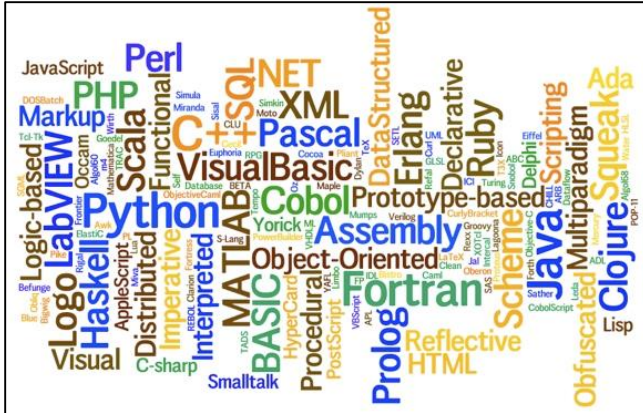


Figura 1 – Nuvem de Palavras

Na figura 2 podemos ver um exemplo que retrata a influência na origem de várias linguagens de programação;

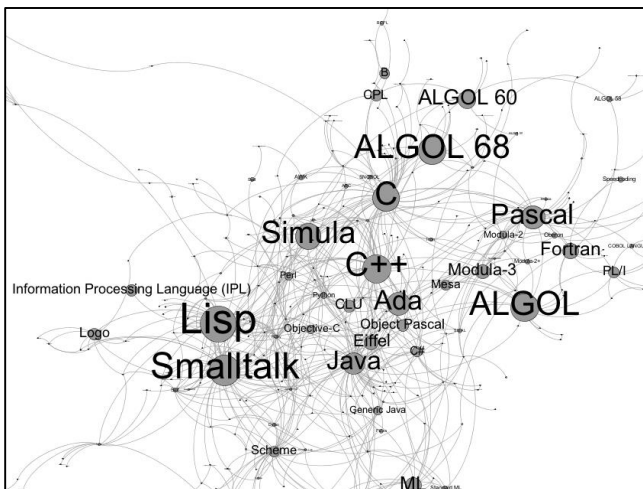


Figura 2 – Rede Neuronal

Fugindo ao domínio de linguagens de programação podemos procurar conteúdo mais focado no programador e nas suas características, no entanto apenas encontramos infogramas maioritariamente sobre web development e outros pouco fidedignos. A figura 3 apresenta um breve infograma que tem como objetivo apresentar algumas das diferenças entre designers e developers.

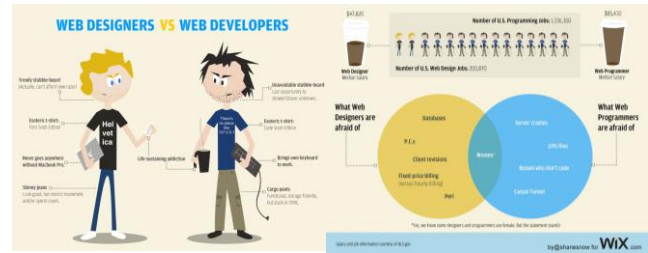


Figura 3 - Infograma

Quanto ao domínio específico do Stack Overflow o trabalho existente é muito menor. Existem alguns estudos feitos por membros da comunidade e ainda algumas VI's, mas que raramente não procuram entender o programador/utilizador do Stack Overflow.

A figura 4 apresenta um screenshot de uma VI que trata várias tags do Stack Overflow. O principal objetivo desta visualização passa por apresentar as principais relações entre tags.

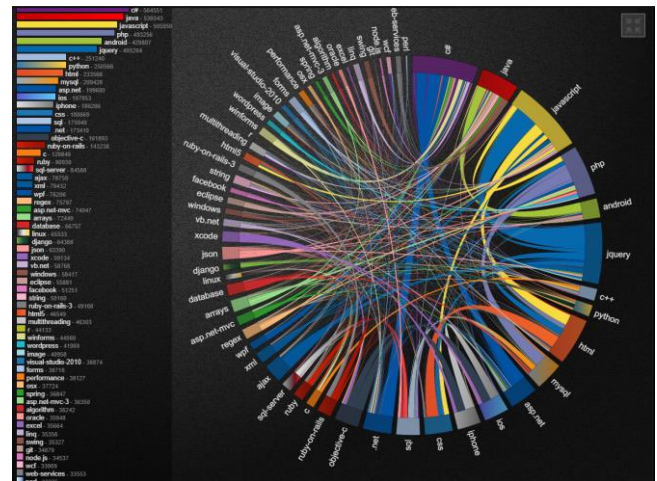


Figura 4 - <http://graphoverflow.com/graphs/stackoverflow-tag-relations.html>

Dito isto não é do conhecimento do grupo nenhuma visualização que trate a informação do Stack Overflow para além das tags. Este facto foi suficiente para o grupo se direccionar no sentido de representar a informação disponível num registo mais espacial e tirar outro tipo de conclusões, como por exemplo: que continente é mais predominante no desenvolvimento de aplicações android.

A DATA

Como explicado na introdução o grupo retirou toda a sua data através do *data explorer* oferecido pelo Stack Exchange. Esta informação está organizada em tabelas e foi necessário algum estudo antes de a extrair através de queries. As figuras 5 e 6 apresentam as duas maiores tabelas na base de dados do Stack Exchange. Através destas e outras tabelas procurámos os dados necessários para construir aquele que seria o nosso dataset; post count, nome da tag, data, região, top users, número de questões e número de respostas.

[illegible]

Figura 5 – User Table

[illegible]

Figura 6 – Post Table

Uma vez compreendida a configuração das tabelas, o grupo iniciou um workshop interno sobre T-SQL entre os três elementos do grupo. O poderoso *data explorer* do Stack Exchange ofereceu a possibilidade de tratar os dados ainda antes da sua extração. Através de queries o grupo recolheu todo o dataset que precisava, a figura 7 mostra a interface onde o grupo escreveu todas as queries necessárias.

Viewing Query

Enter a title for your query

edit description

```

1 SELECT Top 5000 Location, Id AS [User Link], Reputation,
2 Age
3 FROM Users
4 WHERE Location is not null
5 ORDER BY Reputation DESC;
```

Database Schema

Posts	
Id	int
PostTypeId	tinyint
AcceptedAnswerId	int
ParentId	int
CreationDate	datetime
DeletionDate	datetime
Score	int
ViewCount	int
Body	nvarchar(max)

Revisions
488323

Figura 7 – Interface T-SQL

Um dos desafios do projeto relacionava-se com a apresentação da informação organizada por continentes. Acontece que não existe nas tabelas do Stack Exchange nenhuma coluna com essa informação, apenas o local onde o post foi escrito. Para solucionar isto o grupo teve de agrupar primeiro todos os locais por país e só então todos os países por continente. Felizmente procurando pelas queries já existentes na plataforma encontramos uma solução para a primeira agregação de locais. Caso não tivéssemos acesso a esta query o trabalho por registo espacial tornava-se muito mais complicado. Já a segunda agregação foi realizada manualmente, isto porque o volume de países não era excessivo e não chegava às centenas. Com a informação em *.csv* aberta em *Excel* manipulámos o dataset sem hesitações, criando uma nova coluna com o continente para cada país. Ainda removemos países com valores incorretos e corrigimos duplicados com recorrendo um script. Numa fase mais avançada do trabalho tivemos a necessidade de criar uma nova coluna com os ID's de cada país. Estes ID's tinham de cumprir as normas ISO e foram confirmados um a um.

Outro desafio do grupo revelou-se sempre que a data que queríamos extrair tinha mais de 50.000 linhas. A solução baseou-se em escrever várias queries de 50.000 linhas e juntar os vários resultados finais. Apesar de simples, este contra tempo atrapalhou em parte o grupo.

Para correlacionar o número de perguntas e respostas de cada país o grupo derivou uma medida nova. Em Excel e utilizando as colunas de perguntas e respostas criámos a coluna QA. Esta nova medida tem como objetivo apresentar a diferença entre perguntas e respostas de cada país e resume-se em:

$$QA = \frac{\text{respostas} - \text{perguntas}}{\text{respostas} + \text{perguntas}} \times 100$$

Esta formula apenas apresenta valores entre -100 e 100. Países com valores negativos colocam mais perguntas que respostas, enquanto valores positivos revelam países com mais respostas que perguntas.

A VISUALIZAÇÃO

Descrição Geral

A visualização distingue-se em dois modos diferentes selecionáveis através dum dashboard comum.



Figura 8 – Dashboard inicial

Selecionando a *Overall View* o utilizador tem acesso à visão geral sobre as tags no Stack Overflow. Este modo apresenta dois idiomas diferentes; um *colored stacked barchart* e um *colored radarchart*. No *barchart* é possível visualizar o número de tags ao longo do tempo enquanto no *radarchart* é possível visualizar o número de tags por região. A figura 9 apresenta uma visão geral deste modo.

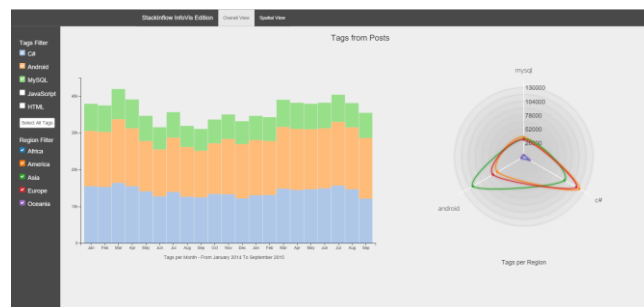


Figura 9 – Overall View

À direita existem várias *checkboxes* coloridas com diferentes cores. A cor de cada caixa de seleção representa

a cor de cada tag no *barchart*. Já as cores em cada caixa de seleção para as regiões representam as cores das regiões no *radarchart*. A figura 10 apresenta em detalhe estas *checkboxes*.

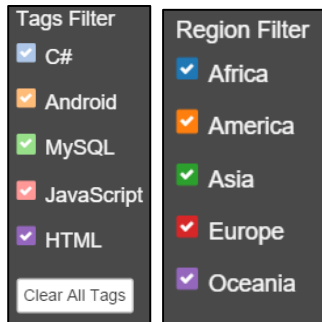


Figura 10 – Checkboxes e cores utilizadas

Selecionar as *checkboxes* adiciona ou remove dados nos charts. Sempre que uma tag é selecionada os valores desta aparecem através duma animação no *barchart*. Selecionar regiões adiciona ou remove os valores associados à região no *radarchart*. A figura 11 e 12 mostram algumas interações com *checkboxes*.

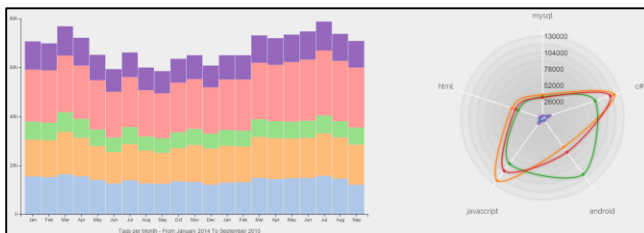


Figura 11 – Charts com todas as checkboxes seleccionadas

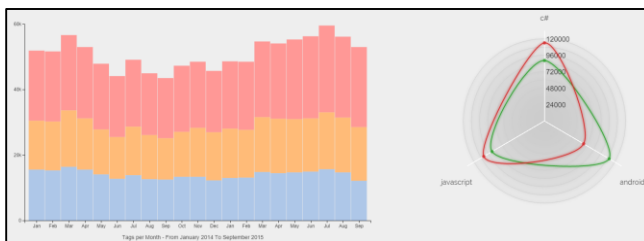


Figura 12 – Charts com algumas checkboxes não seleccionadas

As *checkboxes* de filtro para região afetam também os charts em *Spatial View*, como vai ser demonstrado na secção *Spatial View*.

Sempre que o utilizador faz mouseover sobre uma coluna do *barchart* é devolvido um tooltip com mais informações sobre o nome da tag, o mês e ano da coluna e ainda o número de posts com a tag naquele mês. A cor do tooltip varia dependendo da cor da tag, é possível ver na figura 13 tooltips diferentes que exemplificam este caso.

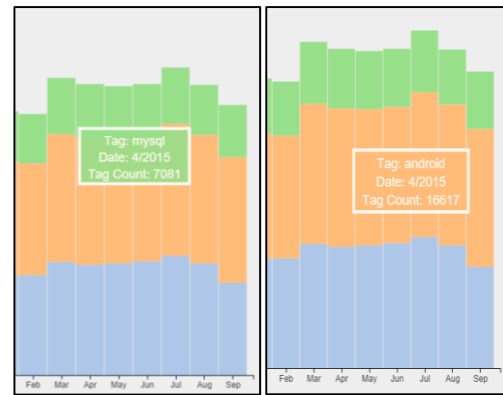


Figura 13 – Mouseover sobre colunas e tooltip

O mesmo acontece no *radarchart*, sempre que o utilizador faz hover num ponto do chart é devolvido um tooltip colorido. A figura 14 mostra dois tipos de tooltip.

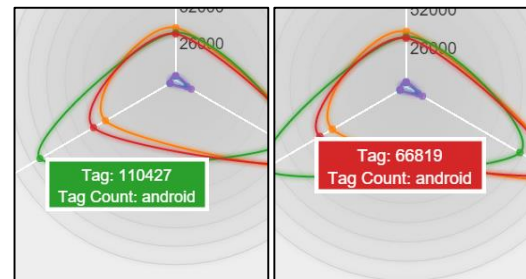


Figura 14 – Mouseover sobre radarchart e tooltip

Caso o utilizador selecione o modo de vista *Spatial View* este irá ser transportado para uma nova visualização semelhante à apresentada na figura 15, com dois gráficos distintos; um *choropleth worldmap* e um *colored scatterplot*.

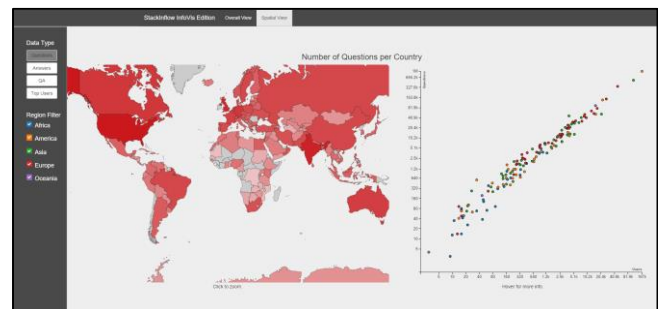


Figura 15 – Spatial View

O dashboard nesta visualização contém as mesmas *checkboxes* para filtrar regiões existentes na *overall view*. Contém ainda quatro botões que trocam a data apresentada. O utilizador pode escolher visualizar “Questions”, “Answers”, “QA” e “Top Users”. A figura 16 mostra em detalhe este botões.



Sempre que o utilizador clicar num destes botões o *worldmap* e o *scatterplot* transfiguram a sua informação de forma animada. Enquanto “Questions” estiver selecionado o *worldmap* apresenta o número de tags pela intensidade da cor vermelha. Os países com mais perguntas apresentam um vermelho mais carregado. Em paralelo o *scatterplot* apresenta o número de perguntas por país no eixo Y e o número de utilizadores por país no eixo dos

X. Assim conseguimos ver muito rapidamente quais os países com maior número de perguntas e concluir através do *scatterplot* se o número de utilizadores influenciou o número de perguntas. Na figura 15 podemos observar tudo o que foi descrito nas últimas linhas.

Se o utilizador optar por selecionar “Answers” o *worldmap* irá apresentar um conjunto de cores azuis. Caso “Top Users” o conjunto de cores será verde, e para representar o “QA” o conjunto de cores vai de vermelho para valores negativos a azul para valores positivos. A escolha de cores não foi aleatória. O vermelho representa algo por resolver enquanto o azul representa algo resolvido e positivo. Na figura 17 encontra-se um exemplo no caso de “Answers” selecionado, na figura 18 um exemplo para “QA” e na figura 19 um exemplo para “Top Users”.

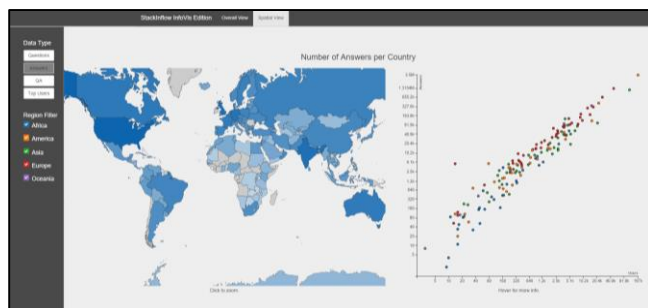


Figura 17 – View for Answers

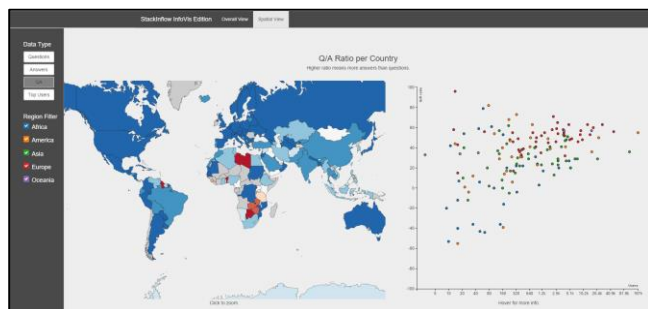


Figura 18 – View for QA

Além da mudança das cores nas figuras 17,18 e 19 podemos também visualizar as alterações no *scatterplot* sempre que selecionamos um dos botões. A maior diferença acontece na vista para “QA” e “Top Users”. O *scatterplot* nestas duas vistas reconfigura a sua escala e eixos de forma a apresentar a data da melhor forma. Em “QA” o eixo Y

começa no valor -100 e acaba em 100, sendo 0 o valor médio deste eixo. Quanto à vista para “Top Users” esta remove todos os países sem top users do domínio, modificando os valores no eixo X e utilizando apenas valores de 1 a 200 no eixo Y. Todas estas reconfigurações no *scatterplot* são feitas utilizando transições, explicando através de animações a diferenças entre cada tipo de dados.

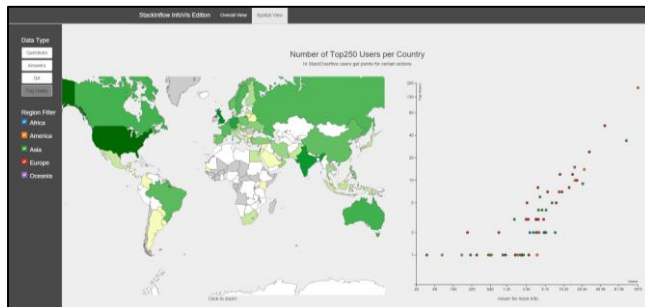


Figura 19 – View for Top Users

Caso o utilizador pretenda apenas visualizar dados referentes a continentes específicos este pode selecionar nas *checkboxes* e filtrar as regiões a apresentar. Estas *checkboxes* preservam as escolhas do utilizador mesmo quando este muda de *Spatial View* para *Overall View*. A figura 20 apresenta um exemplo para “Answers” em que Asia é filtrada. Tanto o *worldmap* como o *scatterplot* deixam de apresentar valores para Asia.

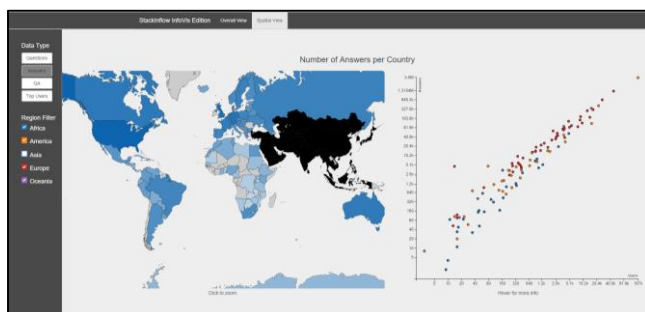


Figura 20 – QA selecionada e África filtrada

A visualização oferece ainda a capacidade de fazer zoom no *worldmap* e mouseover para recolha de mais informação. A figura 21 mostra o que acontece sempre que o utilizador clica no *worldmap*.

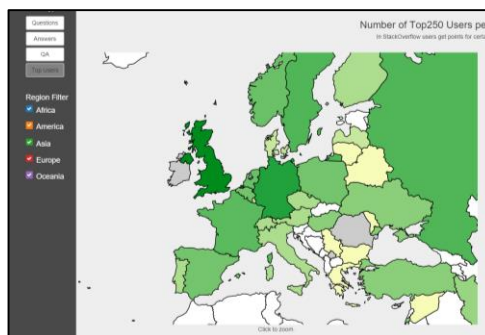


Figura 21 – Zoom sobre a Europa

O tooltip devolvido no mouseover do *worldmap* difere do tooltip devolvido no mouseover do *scatterplot*. A figura 22 mostra os dois tipos de tooltip, primeiro no *worldmap* seguido no *scatterplot*. Sempre que um ponto é mouseovered no *scatterplot* este cresce com uma animação, ficando com maior área e assim mais visível. As cores dos restantes pontos são removidas para acentuar o highlight, perceptível na figura 22.

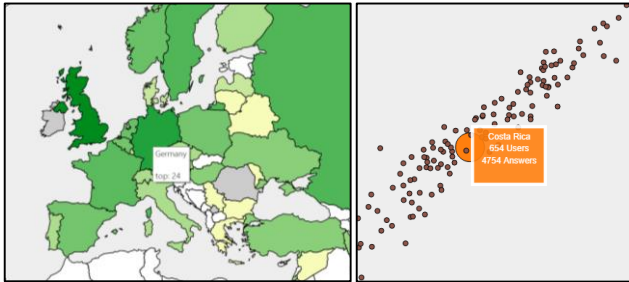


Figura 22 – Mouseover e tooltips

Através desta visualização esperamos que o utilizador consiga tirar diversas conclusões, como quais os países com mais top users e se algum atributo influencia este número, ou que tecnologias são mais utilizadas por região e se essas regiões colocam mais respostas que perguntas.

Rationale

A escolha de idiomas foi realizada desde cedo com o tipo de dados em mente. O estilo de cada chart pouco pesou nas considerações do grupo. Sabíamos o que queríamos apresentar e fomos nessa direção. Esta prática revelou-se um sucesso mais tarde. Na implementação dos protótipos iniciais começamos logo por utilizar o nosso dataset. Acontece que em várias situações temos valores de grande dimensão que implicam a utilização de escalas logarítmicas. Caso tivéssemos optado por outro tipo de idiomas a visualização da informação seria comprometida.

Para apresentar o volume de tags por mês o grupo optou por um *stacked colored barchart*. Este chart funciona por conseguirmos representar três dimensões; número de tags, mês e nome da tag. A utilização de barras é essencial para realizar comparações entre os diferentes valores dos meses. Como explicado nas aulas teóricas, o alinhamento e comprimento são as técnicas mais corretas para comparação de valores, um barchart oferece estas duas características. As cores complementam o chart na medida que um conjunto de cores qualitativas representam facilmente as diferentes tags. Outras opções como conjuntos sequenciais ou divergentes não seriam corretos, as tags são atributos qualitativos.

Para apresentar as tags por região utilizámos um *colored radarchart*. Neste caso o grupo não se baseou tanto nos dados mas sim no que o chart oferecia. A ideia seria apresentar de uma forma rápida as diferenças e características de cada região no que toca a utilização de tags, e assim perceber que tipo de aplicações são desenvolvidas em cada continente. O *radarchart* pode ser

considerado um gerador de figuras, dado um determinado input este devolve uma figura única. Assim com os valores de cada continente conseguimos formar figuras que caracterizam cada continente. O conjunto de cores qualitativo é novamente escolhido para apresentar cada continente como atributo qualitativo. Outros charts poderiam ser utilizados para apresentar tags por região; *bubblecharts* em que a área de cada bolha representa o número de tags e coloridas por região, ou *small multiples* para cada tag ou região. No entanto nenhum outro chart iria caracterizar cada região tão bem.

O *choropleth worldmap* foi pensado juntamente com o *scatterplot*. Estes dois idiomas em conjunto apresentam a reconfiguração do mesmo conjunto de dados. Através do *worldmap* conseguimos ter a visão espacial dos dados enquanto através do *scatterplot* temos a visão das correlações dos dados com o número de utilizadores, conseguindo assim o melhor dos dois mundos. Estes dois charts contêm valores da ordem das dezenas à ordem dos milhões. Esta característica implicou a utilização de uma escala logarítmica que tornou todo o projeto mais desafiante. A visualização permite rodar entre vários tipos de data, e acontece que não utilizam todos a mesma escala. Para apresentar as questões, respostas e top users tivemos de aplicar escala logarítmica, mas só mais tarde nos lembrámos que para o QA a escala teria de ser outra, uma vez que os valores se encontram entre -100 e 100. Foi necessário criar uma escala para cada tipo de dados, isto porque para além dos pontos desenhados sempre que a nova data é selecionada os eixos têm de apresentar a nova informação, o que obrigou a um certo cuidado na escolha do domínio e range em cada caso. Inicialmente tínhamos planeado representar o QA no *worldmap* com uma escala de cores roxa e assim representar esta métrica como um colorblending do vermelho (questões) e respostas (azul). No entanto rapidamente concluímos que este blending não caracterizava a nossa métrica, isto porque seria um conjunto de cores sequencial para um conjunto de dados divergente. O gradiente roxo seria melhor utilizado caso pretendéssemos apresentar o número total de perguntas e repostas de cada país.

Demonstração de Potencial

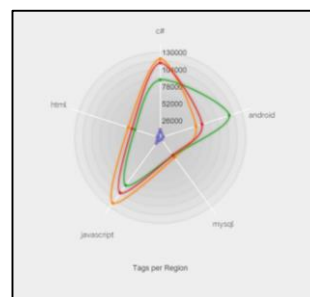


Figura 23 – Ásia a verde

Uma rápida utilização da visualização revela segredos nunca antes descobertos. Visualizando a figura 23 percebemos que existe uma tendência de programadores asiáticos criarem posts para Android. Se considerarmos que alguém só cria um post

em determinada tecnologia se estiver a trabalhar nessa

mesma tecnologia, podemos declarar que na Ásia existe uma predominância de programadores Android.

Também percebemos que existem meses mais prolíferos que outros. Através do *barchart* representado na figura 24 conseguimos ver que nos meses de verão (Agosto, Setembro) existe uma queda de posts.

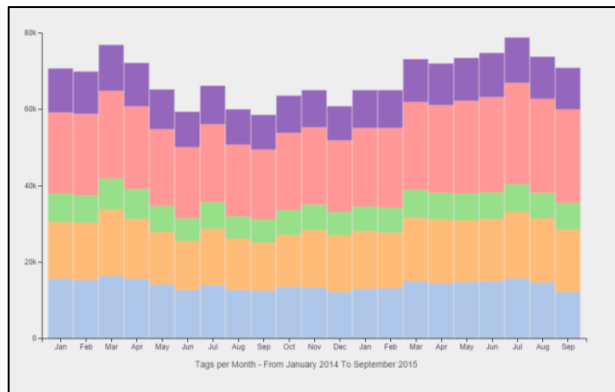


Figura 24 – Volume de posts desde Janeiro 2014 a Setembro 2015

Outras conclusões são também imediatas, como os Estados Unidos da América utilizarem muito mais o site que qualquer outro país. Também é perceptível através da métrica “QA” que os países mais subdesenvolvidos ou de baixa população procuram mais ajuda do que aquela que fornecem.

IMPLEMENTAÇÃO DE DETALHES

Grande parte do nosso trabalho passou por procurar primeiro implementações já existentes e então adaptar às nossas necessidades. Esta prática tem os seus pros e contras. Em sacrifício de alguma flexibilidade ganhamos novas funcionalidades que seriam de difícil implementação. Um exemplo disto é a tentativa de zoom no *scatterplot*. A implementação inicial deste chart trazia cores e pontos com tooltip tal como queríamos. Felizes com estas características optamos por continuar com o código inicial. No entanto mais tarde sentimos a necessidade de fazer zoom no chart. Uma breve pesquisa sobre zoom semântico mostra como é fácil criar esta funcionalidade, mas acontece que o nosso código já existente não aceita este comportamento. Para que tal aconteça o código tem de ser refactorizado várias vezes, não justificando de todo o trabalho. Para contornar este problema fizemos cada ponto aumentar a sua área sempre que existir mouseover. Ainda criámos uma restrição em *javascript* que faz todos os pontos fora do mouseover perderem a sua cor, dando mais ênfase ao ponto selecionado.

Outras situações semelhantes ocorreram; o *worldmap* aceita diferentes escalas para os dados, no entanto a legenda apenas está preparada para receber escalas *quantize*. Entre o compromisso de ter o mapa colorido corretamente ou uma legenda a escolha tombou para a primeira opção.

O plugin de *radarchart* revelou-se complicado na receção de dados. Foi necessário fazer parse em *javascript* de todos os dados lidos pela função *csv* do *d3*. Para construir o *colored stacked barchart* tivemos de implementar de raiz toda a componente de cor e *stacked* com uso a um *barchart* já existente.

O dashboard da visualização começou inicialmente por ser um template *bootstrap*. Eventualmente a visualização tinha ativo o scrolling na página que mais tarde foi removida para melhorar a usabilidade. A animação entre cada modo é totalmente gerada pelo *bootstrap*, no entanto a interação entre idiomas foi criada com uso a *jquery* e *javascript*. A coluna lateral contém várias *checkboxes* que através do seu *id* invocam nos scripts de cada chart condições especiais, conseguindo desta maneira interação entre os vários idiomas. As cores de cada continente são partilhadas em toda a visualização através do uso dum mesmo vetor de cores. O tooltip para mouseover é fornecido pelo *bootstrap*, sendo genérico em todos os idiomas à exceção do *worldmap*.

CONCLUSÃO E TRABALHO FUTURO

O grupo não considera que o projeto esteja ainda concluído. Apesar de conseguirmos com sucesso realizar todas as tarefas a que nos comprometemos, a cada linha de código implementado surgem novas ideias. De facto os 3000 euros e 1 mês para concluir o projeto seriam espremidos ao máximo. Em primeiro lugar gostaríamos de enriquecer os idiomas já existentes com polimentos finais. A interação entre charts enquanto na *Spatial View* não ficou completa sem o mouseover de países no *worldmap* a fazer highlight no *scatterplot*, e vice-versa. Como explicado no capítulo anterior, a capacidade técnica comprometeu o grupo, mas sabemos que com o uso de mais tempo seria uma questão de apenas algumas horas. Da mesma forma o zoom no *scatterplot* e a interação entre *worldmap* e *scatterplot* através de mouseover seria algo essencial, no entanto ficou por implementar, e é sem dúvida o próximo grande passo. Para concluir o projeto atual gostaríamos ainda de adicionar a possibilidade de apresentar no *barchart* os dados com valores percentuais. Isto seria fundamental para a comparação de valores entre diferentes tags. Apesar de ser possível realizar esta tarefa, valores percentuais iriam equilibrar melhor os resultados, reduzindo erros de análise. O alinhamento entre os vários valores para uma dada tag ao longo do tempo seria também um pormenor muito relevante a adicionar.

Dito isto apenas nos resta eventualmente implementar outros idiomas e tornar a experiência de utilização melhor. Seria interessante abordar utilizadores específicos, e poder analisar um a um. Perceber a utilização de tags mais em detalhe seria outro objetivo do grupo. Sem dúvida que tiramos conclusões interessantes com o que já implementámos, mas perceber a fundo o que distingue um programador asiático de um outro europeu seria fascinante.

A que horas trabalha um asiático, e um europeu? Existe semelhanças?

Quanto tempo demora um utilizador a tornar o seu QA positivo?

Um QA positivo em determinadas tags implicam que determinado utilizador é perito nessas mesmas tags?

Estes são alguns exemplos da informação que gostaríamos de extrair. Sem dúvida que a cadeira de VI libertou no grupo a curiosidade e ambição de realizar data mining em tudo o que for possível, e assim ganhar o super poder de argumentar de forma muito mais segura com informação estatística real e de fácil visualização.